Involution Transformer Based U-Net for Landmark Detection in Ultrasound Images for Diagnosis of Infantile DDH

Tianxiang Huang, Jing Shi, Juncheng Li, Jun Wang, *Member, IEEE*, Jun Du, and Jun Shi, *Member, IEEE*

***Abstract*—The B-mode ultrasound based computer-aided diagnosis (CAD) has demonstrated its effectiveness for diagnosis of Developmental Dysplasia of the Hip (DDH) in infants, which can conduct the Graf's method by detecting landmarks in hip ultrasound images. However, it is still necessary to explore more valuable information around these landmarks to enhance feature representation for improving detection performance in the detection model. To this end, a novel Involution Transformer based U-Net (IT-UNet) network is proposed for hip landmark detection. The IT-UNet integrates the efficient involution operation into Transformer to develop an Involution Transformer module (ITM), which consists of an involution attention block and a squeeze-and-excitation involution block. The ITM can capture both the spatial-related information and long-range dependencies from hip ultrasound images to effectively improve feature representation. Moreover, an Involution Downsampling block (IDB) is developed to alleviate the issue of feature loss in the encoder modules, which combines involution and convolution for the purpose of downsampling. The experimental results on two DDH ultrasound datasets indicate that the proposed IT-UNet achieves the best landmark detection performance, indicating its potential applications.**

***Index Terms*—Developmental dysplasia of the hip (DDH), ultrasound images, landmark detection, involution transformer, involution downsampling.**

I. Introduction

DEVELOPMENTAL Dysplasia of the Hip (DDH) is a prevalent yet critical joint disease in infants, resulting in instability and the potential for hip joint dislocation [1], [2]. Accurate diagnosis of DDH is important for the following treatment. Ultrasound imaging is a routine tool for diagnosis of DDH especially for the infants within 6 months [3].

The Graf's method is one of the most commonly used ultrasound examination techniques for DDH, which diagnoses DDH based on the manually measured α and β angles, as illustrated in Fig. 1(a) [4]. However, this determination is highly subjective, depending on sonologists' expertise. Thus, the computer-aided diagnosis (CAD) for DDH has gained its reputation in recent years. Several deep learning (DL) based algorithms have been proposed for this ultrasound-based CAD [5], [6], [7], [8]. These algorithms mainly aim to measure the α and β angles, which can be divided into two categories [5], [6], [7], [8]: the segmentation- and landmark detection-based DL approaches. The former mainly segments the critical anatomical structures for further angle measurement [5], [6], [7], while the latter directly detects the key points in ultrasound images (Fig. 1(b)) [8]. Consequently, as shown in Fig. 1(c), the detected landmarks can form the three corresponding lines that are then used to calculate the α and β angles.

From the viewpoint of data annotation, the landmark detection-based method is more annotation-friendly for the sonologists than the segmentation-based approach, which then has attracted considerable attention for developing the CAD of DDH. However, the quality of ultrasound images is prone to speckle noise [9], which then increases difficulty for detecting the critical landmarks. To address this issue, it is feasible to explore more valuable information around these landmarks, such as the long-range and spatial-related information hidden in the ultrasound images, to enhance feature representation for improving detection performance. For example, Xu et al. [8] indicated the effectiveness of capturing the long-range information to overcome the noise interference for landmark detection in hip ultrasound images. On the other hand, some

Manuscript received 4 December 2023; revised 15 March 2024; accepted 12 April 2024. Date of publication 17 April 2024; date of current version 7 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271298 and Grant 62301306, in part by the Science and Technology Commission of Shanghai Municipality under Grant 23ZR1422200, in part by the 111 Project under Grant D20031, and in part by the Fund of the Cyrus Tang Foundation, and the Fund of the Education Development Foundation of Shanghai Jiao Tong University. *(Corresponding authors: Jing Shi; Jun Shi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by IRB of Shanghai Children's Medical Center Affiliated to Shanghai Jiao Tong University School of Medicine under Application No. SCMCIRB-K2023027-1.

Tianxiang Huang, Juncheng Li, Jun Wang, and Jun Shi are with the Key Laboratory of Specialty Fiber Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: junshi@shu.edu.cn).

Jing Shi and Jun Du are with the Diagnostic Imaging Center, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China (e-mail: shijing@scmc.com.cn).

Digital Object Identifier 10.1109/JBHI.2024.3390241

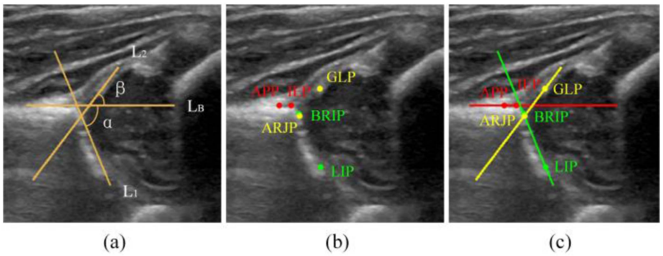

Fig. 1. Illustration of hip ultrasound images for the graf's method. (a) Definition of α and β angles. α is formed by the angle between the base line (L_B) and the bone roof line (L_1), β is created by the intersection of the base line (L_B) and the cartilage roof line (L_2). (b) Six landmarks [8]. 1) Apex point (APP), 2) ilium edge point (IEP), 3) lower limb point (LIP), 4) bony roof incision point (BRIP), 5) acetabular rim junction point (ARJP), 6) glenoid labrum point (GLP). (c) Three critical lines. The red base line (L_B) formed by APP and IEP, the green bone roof line (L_1) formed by LIP and BRIP, the yellow cartilage roof line (L_2) formed by ARJP and GLP.

landmarks are close to each other (e.g., APP, IER, ARJP, and BRIP landmarks as shown in the Fig. 1(b)), which increases difficulties for detection algorithms to accurately detect these landmarks. Thus, the spatial information (i.e., the position of each landmark) is also essential for locating the key points in hip ultrasound images. However, existing algorithms rarely consider the important spatial information. It is still a challenging task to accurately detect the critical anatomical landmarks from the hip ultrasound images.

The convolutional neural network (CNN) with encoder-decoder architecture is commonly applied to the landmark detection task [10]. The U-Net architecture stands out as a representative encoder-decoder architecture, offering numerous advantages [11]. In U-Net, the shallow convolution layers mainly extract texture and edge features, while the deeper convolution layers explore semantic information within the images [12]. A number of works have indicated the superiority of U-Net for image segmentation, reconstruction and landmark detection tasks [13], [14], [15], [16]. However, it still exhibits some limitations for our particular landmark detection task. For example, the pure U-Net architecture based on convolutional layers may not effectively capture global dependencies [12]. Moreover, the translation invariance of convolution operation also makes convolutional layers lacking spatial awareness [17]. In fact, both the long-range interaction and spatial knowledge are essential for accurately localizing key points. On the other hand, U-Net generally adopts maxpooling as the downsampling operation to continually expand the receptive field [12]. However, this maxpooling operation only chooses the maximum value within a local region, thereby sacrificing other valuable information [18]. This may potentially result in the loss of fine-grained details.

Although Xu et al. [8] proposed a relation matrix to capture the long-range information in hip ultrasound images, it may introduce some redundant information that then affects the learning of contextual information. Since Transformer achieves superior performance in capturing long-range dependencies and contextual relationships [19], it can be integrated into the convolution-based U-Net to improve the detection accuracy. In fact, recent studies have tried to build hybrid networks by combining the strengths of CNN and Transformer [14], [15], [20]. These hybrid networks can effectively model global context and capture local features [21]. However, the hybrid models with conventional Transformer architectures generally ignore the modeling of spatial relationships [22]. Thus, it is highly necessary to develop a hybrid U-Net with a spatial-awareness Transformer architecture that specifically for accurate landmark detection.

Furthermore, as a novel atomic operation, involution has a spatial-specific property [23]. Many works have demonstrated that this specifical characteristic can provide the capability to learn spatial information [24], [25], [26], [27]. Different from the traditional convolution operation, the involution generates specific perceptual field weights and adaptively allocates over different positions [24]. Consequently, it has the feasibility to capture the positional information of different landmarks in hip ultrasound images, so as to provide more spatial information for the detection model. Therefore, it is considered that incorporating involution into the Transformer architecture can effectively merge long-range dependency and spatial awareness, so as to enhance the accuracy of landmark detection in hip ultrasound images.

On the other hand, given that the maxpooling results in the loss of important information, it is also important to develop an effective approach for downsampling operation in U-Net. Several previous works have indicated that a convolutional layer with increased stride can effectively replace maxpooling [18]. Due to the superior performance of involution, a novel downsampling module that combines the advantages of the involution and inception is developed in [28]. Inspired by this approach, we believe that it is feasible to develop a new involution-based downsampling method to retain the valuable details around the anatomical landmarks, so as to further improve the detection accuracy.

In this work, a novel Involution Transformer based U-Net network (IT-UNet) is proposed for detecting landmarks from infantile hip ultrasound images. In the encoder-decoder architecture of IT-UNet, a novel Involution-based Transformer Module (ITM) is developed to be embedded in the bottom of U-Net, which can improve the semantic feature representation for landmark detection. Moreover, a new Involution-based Downsampling Block (IDB) is designed in the encoder to perform the downsampling process instead of the traditional maxpooing, which can preserve more detailed information around landmarks. The experimental results indicate the effectiveness of the proposed IT-UNet.

The main contributions of this work are as follows:

1) A novel IT-UNet is proposed to detect landmarks from hip ultrasound images, which can capture and learn more effective feature representation, including both the spatial-related knowledge and long-range dependencies, for improving detection performance.
2) A new ITM is developed by integrating involution into Transformer, which consists of an involution attention block and a squeeze-and-excitation involution block. The

ITM can effectively capture not only the long-range information but also positional information to improve the feature representation.

3) A new IDB is proposed to alleviate the issue of detailed information loss in the encoder module. The IDB innovatively combines involution and convolution for the purpose of downsampling, which can extract more useful features simultaneously.

II. Related Work

A. DL-Based Methods for DDH Diagnosis

In recent years, DL has gained its reputation in the field of ultrasound-based CAD for DDH. Most of these methods focus on developing special segmentation algorithms for the critical anatomical structures to perform the followed angle measurement. For example, Golan et al. [5] implemented a convolutional network with an adversarial component to segment the ilium and acetabular roof, which were subsequently employed to draw lines for calculating the α angle; Hu et al. [7] proposed a multi-task network with Mask R-CNN as backbone, which included a detection and a segmentation branch to mark the four anatomical structures and a landmark detection branch to further measure the two angles from hip ultrasound images; Stamper et al. [29] proposed a lightweight multi-class U-Net network to segment key anatomical structures for DDH screening. All these works have suggested the feasibility of the segmentation-based approaches.

However, the accuracy of angle measurement extremely depends on the performance of segmentation algorithms in these works [30]. Moreover, the segmentation-based methods require professional but laborious annotation, which generally results in the problem of small size samples [31]. The limited training samples then affect the training of DL model. On the contrary, the landmark detection-based DL approaches are simpler and more convenient for annotation. Xu et al. [8] proposed a novel network named Dependency Mining ResNet (DM-ResNet) by a relation matrix, which aimed to combine both short-range and long-range dependencies for landmark detection in hip ultrasound images. This pioneering work indicates the feasibility of landmark detection for calculating α and β angles. However, the global relation matrix inevitably introduces redundant information to the model. Moreover, this model also ignored the valuable spatial information of hip landmarks. Therefore, there is still room for improving the detection performance of hip landmarks.

In this work, we aim to explore both the long-rang dependencies and spatial information to learn more effective feature representation for accurately landmark detection in hip ultrasound images.

B. Involution Network

Involution has the characteristic of spatial-specific, which can be used to explore diverse interactions among spatial locations [23]. Due to this valuable property, involution has been applied to various DL models. For example, Shao et al. [24] proposed a spatial-spectral Involution MLP network for hyperspectral image classification, which utilized involution to extract spatial contextual information in a stable windowed receptive field; Hou et al. [25] developed an Attention-Involution model for visual tracking, which adopted an attention mechanism to generate involution kernels to capture both long-distance and local relations of features. Additionally, involution has also received widespread attention in the field of medical image analysis. For instance, Jain et al. [26] designed a polyp segmentation model by utilizing both the convolution and involution to extract long-range feature dependencies and spatial patterns; Asiri et al. [27] proposed a novel Involution neural network for brain tumor classification, which employed the spatial adaptability of involution to capture intricate features within medical images. All these works have indicated the effectiveness of involution, particularly its spatial awareness.

In this work, we introduce a novel involution-based Transformer module, which seamlessly integrates involution into the Transformer. This structure aims to capture long-range dependencies as well as spatial information, so as to enhance the accuracy of landmarks localization.

C. Downsampling in U-Net

As a commonly used backbone in image segmentation and detection task, U-Net has shown its effectiveness in landmark detection [32], [33], [34]. However, the maxpooling operation in downsampling of U-Net generally suffers from the issue of losing sensitive information, due to compression and aggregation of information [18].This intrinsic property makes U-Net unable to extract detailed information effectively, which is in fact detrimental to detect landmarks. Therefore, various efforts have been made to address this issue. For instance, Stergiou et al. [35] introduced a novel pooling method called SoftPool, which used softmax within a kernel region to better preserve informative features during downsampling process; Kwon et al. [36] used two diagonal elements of downsampling operation instead of maxpooling; Li et al. [37] designed a patch merging refiner module to remove noise and retain the authentic information of feature space. These previous works endeavor to design effective pooing methods to replace maxpooling operation, with the aim of preserving valuable information during downsampling process.

It is worth noting that Xiao et al. [28] combined the advantages of the involution and inception structures by embedding involutions into the maxpooling layer. The involution enables feature extraction within a smaller receptive field, thereby capturing finer-grained features and enhancing perception. However, due to the continued use of maxpooling, the problem of losing information has not been effectively resolved. Therefore, we develop a novel involution-based downsampling block by combining convolution and involution, with the specific aim of preserving fine-grained features.

III. Methodology

As shown in Fig. 2, a novel IT-UNet model is proposed for hip landmark detection from ultrasound images. The IT-UNet is developed based on an encoder-decoder architecture to generate

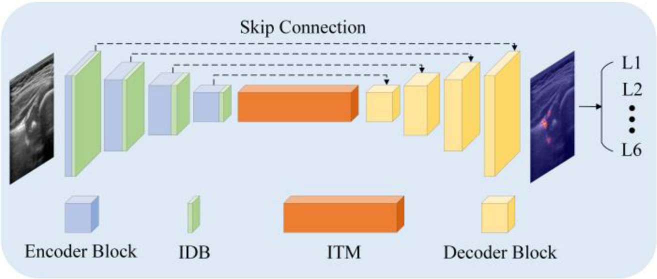

Fig. 2. Overall architecture of IT-UNet, which consists of an involution transformer module (ITM), involution downsampling block (IDB), encoder blocks, decoder blocks, and skip connections. The L1 to L6 represent the predicted coordinates of the six landmarks.

the related heatmaps for further predicting the coordinates of the hip landmarks. The pipeline of IT-UNet mainly includes the following four steps:

1) A hip ultrasound image is first fed into the carefully designed Encoder Block, which can learn valuable feature representation with less detail loss by the developed Involution Downsampling Block (IDB).
2) The extracted feature maps are then fed into the proposed Involution Transformer Module (ITM) to capture the spatial information and long-range dependencies of the ultrasound images.
3) The improved feature representations are subsequently fed into the Decoder Block to restore the feature maps and further generate heatmaps with the help of the skip connection and the upsampling layers.
4) The heatmaps are final used to predict the coordinates of each hip landmark via selecting the position with the maximum heatmap value.

The details of the proposed ITM and IDB are introduced in Sections III-A and III-B, respectively.

A. Involution Transformer Module

To better capture the global spatial dependencies of landmarks, an ITM is proposed to leverage the long-range modeling capabilities of Transformer and the spatial specificity of involution. In particular, the ITM develops a new Involution Attention Block to learn long range semantic context with spatial knowledge. Meanwhile, a Squeeze-and-Excitation Involution Block is designed to fuse hierarchical spatial features with channel-wise information [38].

1) Involution Attention Block: Inspired by the CvT [39], we introduce involution into Transformer and propose the Involution Attention block. As show in Fig. 3(a), Involution Attention contains an Involution Patch Embedding, an Involution Projection, and a Multi-Head Self-Attention (MHSA). The Involution Patch Embedding splits the input images or feature maps into a sequence of patches for information encoding. The Involution Projection operation generates query, key, and value vectors for information transportation. After that, MHSA calculates query, key, and value vectors for modeling long-range feature dependencies.

Before introducing involution, we first define the traditional convolution in existing works. Denote $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C_I}$ as the input feature map, where H, W, and C_I represent the high, width, and channels, respectively. Moreover, the convolution kernel with the size of $K \times K$ is denoted as $\mathcal{F} \in \mathbb{R}^{C_O\times C_I\times K\times K}$, where C_O is the output channels. Thus, output feature map $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times C_O}$ could be calculated by the following formulation:

$$\boldsymbol{Y}_{i,j,o} = \sum_{c=0}^{C_I} \sum_{(u,v)\in\Delta_K} \mathcal{F}_{o,c,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor} \boldsymbol{X}_{i+u,j+v,c} \quad (1)$$

where $o \in [0,\ C_O)$ and Δ_K denotes the neighborhood of center pixel:

$$\Delta_K = \{-\lfloor K/2\rfloor, \ldots, \lfloor K/2\rfloor\} \times \{-\lfloor K/2\rfloor, \ldots, \lfloor K/2\rfloor\} \quad (2)$$

where × indicates Cartesian product here [40].

It is observed that convolution kernel only depends on the numbles of channels and the size of kernel, which makes it spatial-agnostic. However, spatial features show great importance of query (q) and value (v) vectors generation, since q and v represent the queries and representations of positional information. Therefore, we introduce a spatial-specific operator named involution, which has the property of capturing positional information in the spatial domain. Denote $\boldsymbol{X'} \in \mathbb{R}^{H\times W\times C}$ as the input and $C_G = \ C/G$ as the channels contains in a group. Involution kernel is $\mathcal{H} \in \mathbb{R}^{H\times W\times K\times K\times G}$, where H, W, K, G represent high, width, kernel size, and group, respectively. Specifically, an involution kernel that represents a pixel $\boldsymbol{X}'_{i,j} \in \mathbb{R}^C$, where i, j are the coordinates in feature map, is defined as $\mathcal{H}_{i,j,\cdot,\cdot g} \in \mathbb{R}^{K\times K}$, $g \ = \ 1,\ 2, \ldots, G$. Thus, the output feature map is defined as:

$$\boldsymbol{Y}_{i,j,p} = \sum_{(u,v)\in\Delta_K} \mathcal{H}_{i,j,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor,\lceil pG/C\rceil} \boldsymbol{X}'_{i+u,j+v,k} \quad (3)$$

The involution kernel generation function can be defined as $\phi :\ \mathbb{R}^C \mapsto \mathbb{R}^{K\times K\times G}$. Therefore:

$$\mathcal{H}_{i,j} = \ \phi\ (\boldsymbol{X}_{i,j}) = \boldsymbol{W}_1\, \tau\, (\boldsymbol{W}_0 \boldsymbol{X}_{i,j}) \quad (4)$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{\frac{C}{d}\times C}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{(K\times K\times G)\frac{C}{d}}$ denote two linear transformations with a reduction ratio d, and $\tau(\cdot)$ represents the nonlinear transformation that consists of batch normalization and nonlinear activation function.

Noting that involution kernel is sensitive to spatial positions. Therefore, the Involution Projection can generate the q, k and v vectors with spatial awareness:

$$Z^{q/k/v} = FLR(\sigma_R\, (f_{inv}\, (Reshape\, (x)\, , K)) \quad (5)$$

where x is the tokens extracted by Involution Patch Embedding, $f_{inv}(\cdot)$ denotes Involution Projection, K is the kernel size of involution, $\sigma_R(\cdot)$ represents the non-linear activation function ReLu, and $FLR(\cdot)$ is followed by Reshape, Layer Normalization and Flatten.

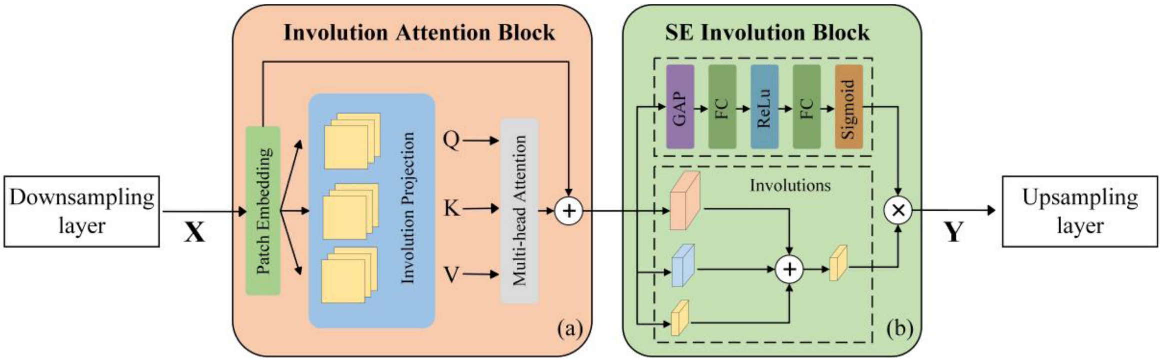

Fig. 3. Structure of involution transformer. (a) Involution attention block. (b) SE involution block.

The generated q, k and v vectors are then fed into MHSA. The total computation of MHSA can be formulated as:

$$MHSA\left(Z^{q/k/v}\right) = softmax\left(\frac{Z^q Z^{k^T}}{\sqrt{d_k}}\right) Z^v \tag{6}$$

Therefore, the output of the Involution Attention block is given as:

$$x' = Z^{q/k/v} + MHSA\left(Z^{q/k/v}\right) \tag{7}$$

2) Squeeze-and-Excitation Involution Block: According to (7), we get features by Involution Attention block. In traditional Vision Transformer, these features will be fed into a Feed Forward Network (FFN) with the residual structure to further perform feature transformation and enhance the nonlinearity. FFN commonly comprises Batch Normalization and Multi-layer Perceptron (MLP). However, the MLP globally operates on all token maps but ignores hierarchical learning of vision representations [41]. Thus, as shown in Fig. 3(b), we design a Squeeze-and-Excitation (SE) Involution block to fuse hierarchical spatial features and channel-wise information.

We first generate hierarchical feature representations by different kernel sizes of involutions. This process could be formulated as:

$$x_i' = D\left(\sigma_G\left(f_{inv}\left(x'\right), k_i\right)\right),\ i = 1,\ 2,\ 3 \tag{8}$$

where $\sigma_G(\cdot)$ denotes the non-linear activation function GeLu, and $D(\cdot)$ is dropout operation that helps enhance robustness to different features. We set three sizes of involution kernel $k_1 = 3,\ k_2 = 5$, and $k_3 = 7$ to extract hierarchical spatial features. The output of this branch could be formulated as :

$$x_{inv} = \sigma_G\left(f_{inv}\left(\sum_{i=1}^{3} x_i'\right),\ k_1\right) \tag{9}$$

Besides, we further propose a SE branch to supplementary the channel information:

$$SE\ (x') = Sigmoid\left(\boldsymbol{W}_{f_1}\sigma_R\left(\boldsymbol{W}_{f_2}GAP\left(x'\right)\right)\right) \tag{10}$$

where $GAP(\cdot)$ denotes the global average pooling, $\sigma_R(\cdot)$ represents ReLu activation function, $Sigmoid(\cdot)$ represents sigmoid function, and $\boldsymbol{W}_{f_1} \in \mathbb{R}^{\frac{c}{r}\times c}$ and $\boldsymbol{W}_{f_2} \in \mathbb{R}^{c\times\frac{c}{r}}$ refer to the two full-connected (FC) layers with the dimensionality reduction ratio r. Therefore, the output of SE Involution is given as:

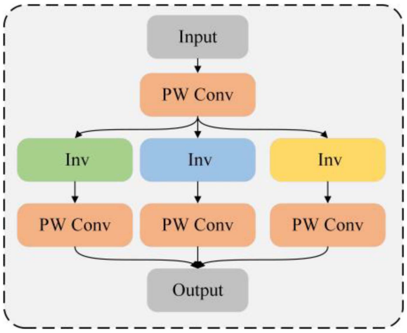

Fig. 4. Structure of involution downsampling. The PW conv is the point-wise convolution and the three invs represent the involutions with three different kernel sizes.

$$\boldsymbol{Y} = SE\left(x'\right) * x_{inv} \tag{11}$$

where $*$ refers to channel-wise multiplication.

By the carefully designed Involution Attention block and SE Involution block, ITM can effectually capture global semantic information with spatial awareness, which then could help precisely detect the landmarks within critical anatomical structures.

B. Involution Downsampling Block

The encoder-decoder architecture is effective for extracting the high-level features (i.e., semantic information) from images. To reduce computation and improve efficiency, dimensionality reduction is the key to this architecture. However, this process inevitably suffers from losing features that can interfere with the downstream task. Different from the traditional maxpooling operation [42], we designed a new Encoder Block in the encoder. The Encoder Block consists of two convolutional layers and an IDB (Fig. 4), which is designed for learning powerful feature representation with less information loss.

In the encoder block, the IDB reduces the dimensionality via a feature extraction process by involutions and convolutions. Denote $\boldsymbol{x}_f \in \mathbb{R}^{h\times w\times c}$ as the input feature map, where h, w, c are the high, width and channels. We first utilize a Point-Wise convolution to reduce the dimensionality of the input. Then, the obtained feature map $\boldsymbol{x}_{f_0} \in \mathbb{R}^{h\times w\times \frac{c}{s}}$ is reduced by half by the involutions with stride of 2. Besides, we introduce three different kernel sizes in involution layers to extract rich information and then obtain three feature maps $\boldsymbol{x}_{f_{1,\ 2,\ 3}} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times\frac{c}{s}}$. The subsequent Point-Wise convolutions are used to restore the number of channels. Finally, the output of IDB is formulated as:

$$\boldsymbol{y}_f = \sum_{j=1}^{3} \left(\sigma_G\left(\boldsymbol{x}_{f'_j}\right)\right) \tag{12}$$

where $\boldsymbol{x}_{f'_j} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c}$, $j = 1,\ 2,\ 3$ are the restored feature maps.

In this way, we redefine downsampling process as feature extraction by involutions and convolutions. Compared to simply selecting the maximum value from a local region, feature mapping operation projects valuable information across overlapping areas. Thus, valuable information in the ultrasound images can be preserved. That is, our proposed IDB can significantly alleviate the problem of losing sensitive features and allow the model to retain detailed information around the landmarks during downsampling process.

IV. Experiments

A. Dataset and Preprocessing

Two DDH ultrasound datasets were used to evaluate the proposed IT-UNet algorithm in this work. The first dataset was collected from the Shanghai Children's Medical Center (SCMC), including 700 hip ultrasound images from 413 infants. Specifically, 500 images of them (named SCMC DDH Dataset A) were scanned by the LOGIQ E9 ultrasound scanner (GE HealthCare, Milwaukee, WI) with an 8.4 MHz linear-array probe between June 2022 and August 2022. Moreover, the images in SCMC DDH Dataset A were scanned with the same dynamic range of 69dB, and the gain values were set within the range of 4.0dB to 6.0dB. The other 200 images (named SCMC DDH Dataset B) were scanned by another ultrasound device (SIEMENS OXANA 2, Inc., Chicago, IL, USA) with an 9MHz liner-array probe between January 2023 to October 2023. In SCMC DDH Dataset B, the images were scanned with the dynamic range of 50dB, and the gain values were ranging from 0dB to 10.0dB. This study was approved by the Research Ethics Board of Shanghai Children's Medical Center, and informed consent was signed by all guardians of the infants. All landmarks were marked by experienced sonologists.

The second APCH DDH Dataset in [8] comprised 1769 hip ultrasound images, which were collected from the Anhui Provincial Children's Hospital between December 2018 to November 2019. These images were scanned by a Philips EPIQ 5 ultrasound system. The landmarks were labeled and cross-validated by four professional doctors who have engaged in DDH diagnosis for more than five years.

There were three sizes of image resolution in the SMMC DDH Dataset, including 368×390, 440×480, and 480×480 pixels. Moreover, all of images in the APCH DDH Dataset had the resolution of 445×715 pixels.

B. Experimental Settings

To evaluate the performance of our proposed IT-UNet, we selected the following eight representative algorithms for comparation, including U-Net [12], DM-ResNet [8], HRNet [43], UNet++ [13], TransUNet [14], FARNet [16], DA-TransUNet [44], and SCUNet++ [45]:

1) U-Net [12]: The classical encoder-decoder architecture U-Net was adopted for landmark detection, which was the baseline network in our experiment.
2) DM-ResNet [8]: This model was specially proposed for hip landmark detection task, which adopted a simple ResNet as the backbone and presented a novel dependency mining module to enhance features.
3) HRNet [43]: This model was a deep convolutional neural network for key point detection, which repeatedly exchanged the information by parallel connecting the high-to-low resolution convolutions.
4) UNet++ [13]: It was a deeply-supervised encoder-decoder architecture, which connected sub-networks through a series of nested, dense skip pathways.
5) TransUNet [14]: It served Transformers as strong encoders and used U-Net to recover localized spatial information that enhanced finer details.
6) FARNet [16]: FARNet was a novel encoder-decoder architecture for anatomic landmark detection, which fused multi-scale features from the encoder and achieved high-resolution heatmap regression.
7) DA-TransUNet [44]: It was a state-of-the-art (SOTA) U-shape architecture, which utilized the Transformers and dual attention blocks to combine not only global and local features but also image-specific positional and channel features.
8) SCUNet++ [45]: It was another SOTA network with multiple fused dense skip connections between the encoder and decoder, which aimed to fuse features of different scales and compensate for the spatial information loss caused by downsampling.

We also conduct the following ablation experiment on the SCMC DDH Dataset to further verify the effectiveness of the ITM and IDB:

1) U-Net [12]: The original U-Net was adopted for landmark detection, without any proposed modules.
2) CvT-UNet: This variant integrated the Convolution Transformer (CvT) into U-Net, which aimed to compare the effectiveness of the developed Involution Transformer.
3) CvD-UNet: This variant utilized convolutions with stride 2 for downsampling (CvD) in U-Net. It aimed to verify the effectiveness of our proposed IDB.
4) IvID-UNet: This variant employed Involution with Inception Downsampling (IvID) block to replace maxpooling in U-Net, which aimed to evaluate the effectiveness of the proposed IDB again.
5) IT-UNet w/o IDB: This variant only applied the proposed ITM in U-Net architecture, which captured both long-range dependencies and spatial knowledge to locate the anatomical landmarks in the hip ultrasound images.

6) IT-UNet w/o ITM: This variant only replaced the traditional maxpooling with IDB in U-Net. It aimed to maintain fine-grained features during downsampling process.

C. Evaluation Metrics

We conducted the five-fold cross-validation strategy to evaluate all the algorithms. All the results were presented in the format of mean ± SD (standard deviation).

Both mean radial error (MRE) and successful detection rate (SDR) were used as the evaluation metrics to verify the performance of landmark detection. The MRE is defined as:

$$MRE_n = \left|L_n^{pred} - L_n^{gt}\right|^2 \tag{13}$$

where $L_n^{pred} \in (x_n^{pred}, y_n^{pred})$ and $L_n^{gt} \in (x_n^{gt}, y_n^{gt})$ represent the $n-th$ prediction and ground truth landmarks. Thus, (13) could further convert into:

$$MRE_n = \left(x_n^{pred} - x_n^{gt}\right)^2 + \left(y_n^{pred} - y_n^{gt}\right)^2 \tag{14}$$

Notably, MRE stands for radial error between the predicted landmark and the ground truth landmark. A smaller MRE value denotes a more precise detection. We also used the SDR as formulated:

$$SDR_{dist} = \#\left\{m : MRE_n \leq dist\right\}/M \times 100\% \tag{15}$$

where $\#$ is used as the count symbol, m denotes the number of landmarks that $MRE_n \leq dist$, M represents the total number of landmarks in images, and $dist$ is the scope of successful detection.

SDR is employed to evaluate the distribution of MRE, with the value indicating the reliability of the landmark detection. In this work, we set $dist$ into 0.5 mm, 1.0 mm, 1.5 mm, respectively.

D. Implementation Details

During the training stage, the input hip ultrasound image was resized to 256 × 256. Meanwhile, the size of mini-batch was set to 2. We set the hyperparameter σ to 10, which determined the Gaussian distribution while generating ground truth heatmaps. The Adam optimizer was utilized for network optimization with a learning rate of 0.0001, and the model was trained for 300 epochs. In addition, the training loss of IT-UNet converged at approximately 250 epochs. All algorithms were implemented by PyTorch with a GTX 2080TI GPU.

V. Experimental Results

A. Results of Comparation Experiments

Fig. 5 shows the visualization results of different landmark detection algorithms on the SCMC DDH Dataset and the APCH DDH Dataset. The red dots represent the ground truth landmarks, the green dots show the detected results by different algorithms, and the yellow lines between the red dots and green dots denote the detected errors. It can be found that the proposed IT-UNet achieves the best detection accuracy, since the predicted landmarks are more closely match the ground truth locations. It is worth noting that despite the differences of the hip ultrasound images, the IT-UNet also achieves the best detection accuracy.

Table I gives the quantitative MRE results of different algorithms on the SCMC DDH Dataset, with the L1 to L6 as the six anatomical landmarks (e.g., APP to GLP in Fig. 1(b)) of hip images. The proposed IT-UNet achieves the best detection results for almost all landmarks except the LIP and BRLP, and gets the best average MRE of 0.4494±0.0155 mm. Compared to other comparison algorithms, it decreases at least 0.0188 mm (about 4.02%). Moreover, compared with the DM-ResNet that is specially designed for the same hip landmark detection task in Reference [8], the average MRE of IT-UNet reduces 0.0367 mm (approximately 7.55%). All these superior results demonstrate the effectiveness of our proposed IT-UNet in accurately localizing landmarks from hip ultrasound images.

Table II further gives the comparison results of SDR on the SCMC DDH Dataset. It can be observed that the proposed IT-UNet again outperforms all the compared algorithms with three best SDR values of 71.19±1.76%, 93.45±1.07%, and 97.31±0.56%, respectively. Moreover, compared to the typical DM-ResNet algorithm, our model improves 2.55%, 2.31%, and 1.02%, respectively, on the corresponding 0.5 mm, 1.0 mm and 1.5 mm. These results demonstrate that the majority of predicted landmarks by IT-UNet are in close to the ground truth landmarks, thereby indicating the effectiveness of the proposed IT-UNet.

Table III shows the quantitative MRE results of different algorithms on the APCH DDH Dataset. Our IT-UNet again outperforms all the compared algorithms for detecting the hip landmarks. The IT-UNet achieves the best average MRE of 0.4282±0.0206 mm, which decreases at least 0.0124 mm (about 2.81%) over all other algorithms. Furthermore, when compared to the SOTA algorithms, such as FARNet, DA-TransUNet, and SCUNet++, the proposed IT-UNet still demonstrates superior performance with a reduction of the average MRE by 0.1888 mm, 0.0144 mm, and 0.0124 mm, respectively.

Table IV presents the three SDR results on the APCH DDH Dataset. It can be found that the proposed IT-UNet achieves the highest scores on the SDR at 0.5 mm and 1.0 mm with values of 72.19±1.60% and 94.25±0.43%, respectively, and improves at least 1.06% and 0.40%, respectively, compared to other algorithms. Moreover, the IT-UNet gets the second-highest score on the SDR at 1.5 mm, and is only surpassed by the FARNet with a decrease of 0.15%.

B. Results of Ablation Experiments

Fig. 6 presents the visual comparison of ablation study. Notably, the IT-UNet exhibits the most superior visual detection performance. Moreover, it is observed that some landmarks (green dots) predicted by IT-UNet w/o ITM or IT-UNet w/o IDB deviate from the ground truth landmarks (red dots). Visual results indicate that both the ITM and IDB are essential for our IT-UNet.

Table V shows the quantitative MRE results of ablation study on the SCMC DDH Dataset. In comparison to the IT-UNet, the IT-UNet w/o ITM shows a decline in performance with an increase of 0.0198 mm (about 4.22%) on the average MRE, while the IT-UNet w/o IDB exhibits a growth of 0.0121 mm (about 2.62%). These results indicate the importance of both the

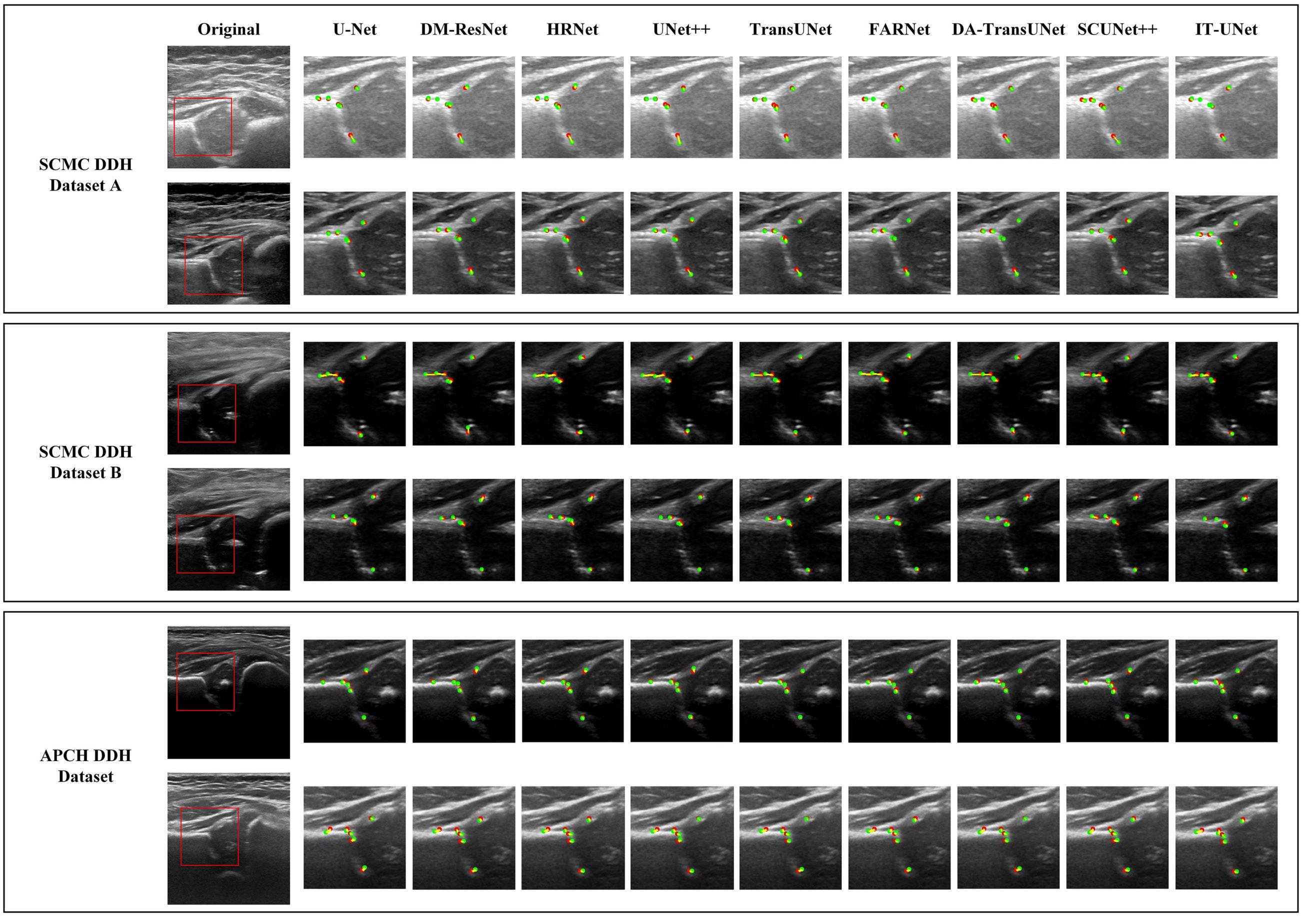

Fig. 5. Visualization results of landmark detection by IT-UNet and other comparison algorithms. The red box in the original hip image represents the area including critical anatomical structures. The red dot represents ground truth landmark while the green dot is the predicted landmark by DL models. The yellow line between the red dot and green dot denotes the detecting errors of DL models.

TABLE I
Quantitative Results of Different Algorithms on the SCMC DDH Dataset With MRE (Unit: mm)

Algorithm	MRE (mm) ↓						
	L_1(APP)	L_2(IEP)	L_3(LIP)	L_4(BRIP)	L_5(ARJP)	L_6(GLP)	Avg
U-Net [12]	0.5264±0.0425	0.4938±0.0461	0.7330±0.0452	0.3794±0.0199	0.4007±0.0255	0.4204±0.0164	0.4923±0.0255
DM-ResNet [8]	0.5385±0.0362	0.5216±0.0578	0.6954±0.0530	0.3620±0.0114	0.3757±0.0258	0.4233±0.0426	0.4861±0.0262
HRNet [43]	0.5287±0.0531	0.5040±0.0597	0.7196±0.0375	0.4242±0.0853	0.4774±0.1188	0.4788±0.0928	0.5221±0.0607
UNet++ [13]	0.5098±0.0341	0.4999±0.0457	0.6741±0.0263	0.3642±0.0136	0.3851±0.0095	0.4034±0.0098	0.4728±0.0172
TransUNet [14]	0.5776±0.0301	0.5303±0.0399	0.7264±0.0715	0.3770±0.0128	0.3954±0.0195	0.4268±0.0225	0.5056±0.0258
FARNet [16]	0.5419±0.0284	0.5118±0.0502	**0.6544±0.0273**	**0.3430±0.0135**	0.3705±0.0110	0.4022±0.0286	0.4706±0.0172
DA-TransUNet [44]	0.5063±0.0266	0.5068±0.0503	0.6675±0.0300	0.3589±0.0158	0.3748±0.0124	0.3951±0.0299	0.4682±0.0213
SCUNet++ [45]	0.5078±0.0303	0.5029±0.0455	0.6823±0.0462	0.3749±0.0139	0.3871±0.0198	0.3902±0.0292	0.4742±0.0216
IT-UNet (Ours)	**0.4830±0.0402**	**0.4603±0.0522**	0.6590±0.0305	0.3471±0.0143	**0.3688±0.0093**	**0.3779±0.0154**	**0.4494±0.0155**

The bold values represent the best results.

proposed ITM and the IDB. Specifically, the removal of ITM in IT-UNet results in an obvious decline in the MRE metric compared with the IT-UNet, because the IT-UNet w/o ITM cannot effectively capture and learn spatial and global information in the hip ultrasound images without the ITM. Similarly, the decreased performance of IT-UNet w/o IDB also demonstrates that this variant suffers from the issue of feature loss during the downsampling process in U-Net. Besides, compared to the CvT-UNet that only integrates the Convolution Transformer (CvT) into U-Net, the IT-UNet w/o IDB that still has ITM shows a reduction of 0.0154 mm on the average MRE (approximately 3.23%). It indicates that although the CvT can capture the long-range information from hip ultrasound images, the proposed ITM can learn additional spatial information besides the long-range dependencies, so as to further enhance feature representation. On the other hand, the CvD-UNet utilizes convolutions with

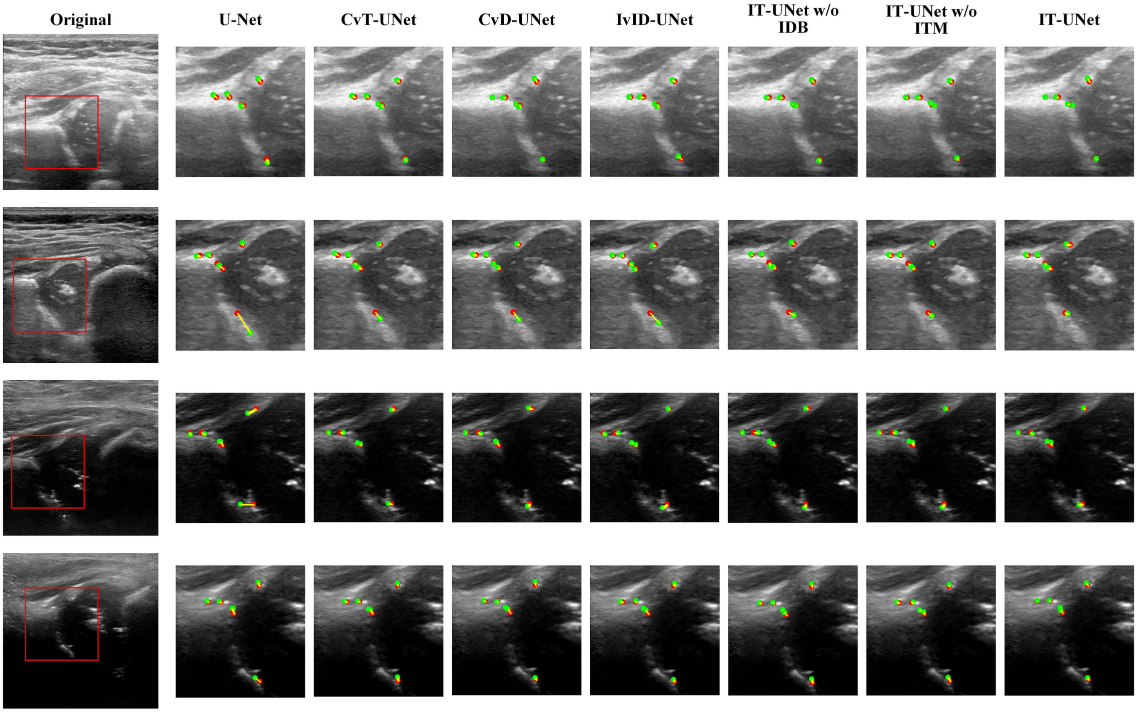

Fig. 6. Visualization results of landmark detection by IT-UNet and the ablation variants. The red box in the original hip image represents the area including critical anatomical structures. The red dot represents ground truth landmark while the green dot represents predicted landmark by DL models. The yellow line between the red dot and green dot denotes the detection errors of DL models.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS ON THE SCMC DDH DATASET WITH SDR (UNIT: %)

Algorithm	SDR (%) ↑		
	0.5 mm	1.0 mm	1.5 mm
U-Net [12]	67.12±1.62	92.48±0.83	97.14±0.71
DM-ResNet [8]	68.64±2.13	91.14±1.41	96.29±0.59
HRNet [43]	66.55±2.64	91.33±1.12	96.29±0.85
UNet++ [13]	68.12±1.07	92.62±0.95	97.31±0.58
TransUNet [14]	67.57±1.25	91.79±0.98	96.45±0.62
FARNet [16]	68.29±0.65	92.62±0.89	97.12±0.71
DA-TransUNet [44]	69.60±0.92	92.17±0.61	96.71±0.84
SCUNet++ [45]	69.43±2.82	92.74±1.49	96.52±0.49
IT-UNet (Ours)	**71.19±1.76**	**93.45±1.07**	**97.31±0.56**

The bold values represent the best results.

stride 2 for downsampling in U-Net, and the IvID-UNet adopts the Involution with Inception Downsampling block to replace maxpooling in U-Net. It can be found the IT-UNet w/o ITM that still has IDB decreases at least 0.0126 mm on the average MRE compared to both IvID-UNet and CvD-UNet, which demonstrates the effectiveness of the developed IDB in preserving valuable details during the downsampling process.

Table VI further presents the results of ablation study on the SDR values. It is notable that after removing the ITM or IDB from the IT-UNet, all the three SDR values of the IT-UNet w/o ITM and the IT-UNet w/o IDB decline compared with the IT-UNet, suggesting the importance of both ITM and IDB. In particular, after removing the proposed ITM, there is a significant decrease in the value of SDR at 0.5 mm, with a reduction of 1.74%. It again indicates the effectiveness of the ITM in capturing both the spatial and long-range information. While compared with the CvT-UNet, the IT-UNet w/o IDB still achieves superior performance on all three SDR values. Moreover, the IT-UNet w/o ITM also outperforms both CvD-UNet and IvID-UNet. These results indicate the same conclusions as mentioned above.

C. Generalization Analysis

To further evaluate the generalization of the proposed IT-UNet, we used the 500 ultrasound images in the SCMC DDH Dataset A as the training set, which were scanned by the LOGIQ E9 ultrasound device, and the other 200 images in the SCMC DDH Dataset B were utilized as the testing set, which were scanned by another ultrasound device (SIEMENS OXANA 2).

Table VII shows the quantitative results on the two metrics. It can be found that the IT-UNet again achieves the best detection results for almost all landmarks except the LIP, and gets the best average MRE of 0.6479 mm. For the three SDR metrics, the IT-UNet again outperforms all compared algorithms, achieving the values of 55.92%, 84.25%, and 91.58%, on the corresponding 0.5 mm, 1.0 mm, and 1.5 mm. It improves at least 2.92%, 2.25%, and 0.36%, respectively, on the corresponding metrics. All these experimental results demonstrate the effectiveness of the proposed IT-UNet, which has superior generalization to all the comparison algorithms.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS ON THE APCH DDH DATASET WITH MRE (UNIT: MM)

Algorithm	MRE (mm) ↓						
	L_1(APP)	L_2(IEP)	L_3(LIP)	L_4(BRIP)	L_5(ARJP)	L_6(GLP)	Avg
U-Net [12]	0.5050±0.0213	0.4201±0.0093	0.4037±0.1226	0.6316±0.0406	0.4934±0.0368	0.3425±0.0071	0.4661±0.0266
DM-ResNet [8]	0.5040±0.0319	0.3978±0.0221	0.4009±0.1099	0.6346±0.0374	0.4657±0.0337	0.3375±0.0174	0.4567±0.0253
HRNet [43]	0.4743±0.0254	0.4100±0.0228	0.3763±0.1284	0.6580±0.0367	0.4711±0.0450	0.3438±0.0088	0.4556±0.0323
UNet++ [13]	0.4508±0.0129	0.4003±0.0129	0.3900±0.1294	0.6336±0.0441	0.4725±0.0410	0.3489±0.0256	0.4493±0.0297
TransUNet [14]	0.5558±0.0140	0.4062±0.0144	0.3839±0.1097	0.6394±0.0418	0.4855±0.0558	0.3400±0.0071	0.4685±0.0261
FARNet [16]	0.4686±0.0145	0.4077±0.0231	0.3790±0.1090	0.6208±0.0319	0.4712±0.0252	0.3343±0.0076	0.4470±0.0193
DA-TransUNet [44]	0.4742±0.0123	0.3980±0.0130	0.3758±0.1197	0.6185±0.0385	0.4620±0.0386	0.3269±0.0053	0.4426±0.0270
SCUNet++ [45]	0.4631±0.0245	0.4010±0.0117	0.3779±0.1181	0.6180±0.0356	0.4444±0.0176	0.3391±0.0064	0.4406±0.0243
IT-UNet (Ours)	**0.4458±0.0168**	**0.3947±0.0273**	**0.3619±0.1034**	**0.6057±0.0439**	**0.4399±0.0245**	**0.3213±0.0179**	**0.4282±0.0206**

The bold values represent the best results.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS ON THE APCH DDH DATASET WITH SDR (UNIT: %)

Algorithm	SDR (%) ↑		
	0.5 mm	1.0 mm	1.5 mm
U-Net [12]	67.72±0.92	93.56±0.35	98.10±0.21
DM-ResNet [8]	70.24±0.62	93.16±0.39	97.69±0.41
HRNet [43]	69.31±1.47	93.21±0.58	97.92±0.23
UNet++ [13]	70.12±1.81	93.73±0.45	98.05±0.22
TransUNet [14]	69.25±1.32	92.42±0.31	97.44±0.19
FARNet [16]	70.22±1.18	93.85±0.33	**98.29±0.18**
DA-TransUNet [44]	71.13±1.37	93.77±0.46	98.12±0.18
SCUNet++ [45]	70.08±0.66	93.80±0.33	98.09±0.25
IT-UNet (Ours)	**72.19±1.60**	**94.25±0.43**	98.14±0.24

The bold values represent the best results.

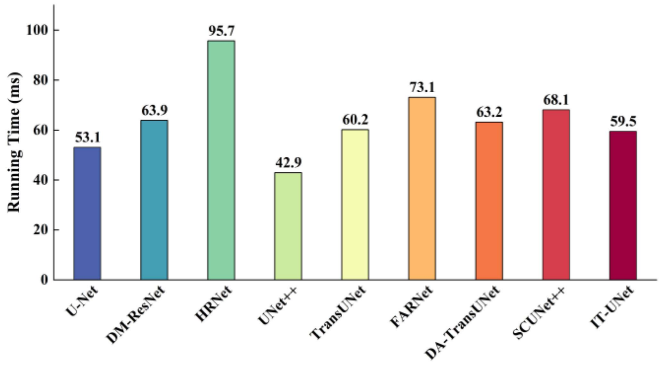

Fig. 7. Average running time of different algorithms to detect the hip landmarks from one ultrasound image.

D. Computational Complexity and Running Time

Fig. 7 shows the average running time of different algorithms to predict one ultrasound image during the testing stage. It can be found that the proposed IT-UNet costs only 59.5ms to detect the hip landmarks from a signal ultrasound image, which is located at the middle level among all algorithms.

Table VIII further gives the model parameters (Params) and floating point operations (FLOPs) of different algorithms for hip landmark detection. It is observed that the IT-UNet has a slightly higher computational cost, with the values of 181.999M Params and 94.411G FLOPs. However, as shown in Fig. 7, it still has an acceptable running time during the testing stage.

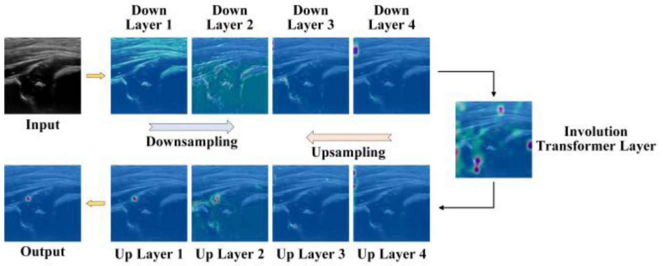

Fig. 8. CAMs of tracking the GLP hip landmark from each layer in the proposed IT-UNet.

E. Visualization of CAMs in IT-UNet

Fig. 8 further illustrates a series of class activation maps (CAMs) [46], which are obtained by tracking one hip landmark (GLP is chosen in Fig. 8) from each layer in the proposed IT-UNet. The CAMs can allow us to understand which parts of the input image the model focuses on when making predictions. It is worth noting that the IT-UNet emphasizes texture information in the shallow down layers, such as Down Layer 1 and Down Layer 2. As the network progresses, the highlighted areas in the CAMs gradually narrow down, indicating a shift in focus towards more localized information. Notably, following the integration of the developed Involution Transformer layer, our IT-UNet exhibits a broader focus, capturing additional positional and global information within the hip image. This intriguing observation underscores the effectiveness of the Involution Transformer module in capturing both positional and global context. During the upsampling process, the IT-UNet progressively shifts its attention towards the neighborhood of the GLP landmark. The final CAM vividly illustrates our IT-UNet's concentration on the neighborhood of the GLP landmark, which serves as a crucial cue for predicting the coordinates of the hip landmark. This comprehensive visualization of the CAMs offers a clear window into the decision-making process

TABLE V
QUANTITATIVE RESULTS OF ABLATION STUDY ON THE SCMC DDH DATASET WITH MRE (UNIT: MM)

Algorithm	MRE (mm) ↓						
	L_1(APP)	L_2(IEP)	L_3(LIP)	L_4(BRIP)	L_5(ARJP)	L_6(GLP)	Avg
U-Net	0.5264±0.0425	0.4938±0.0461	0.7330±0.0452	0.3794±0.0199	0.4007±0.0255	0.4204±0.0164	0.4923±0.0255
CvT-UNet	0.5258±0.0623	0.4999±0.0618	0.6872±0.0383	0.3590±0.0140	0.3899±0.0065	0.3998±0.0206	0.4769±0.0274
CvD-UNet	0.5266±0.0369	0.4915±0.0438	0.6914±0.0253	0.3760±0.0202	0.4002±0.0153	0.4053±0.0249	0.4818±0.0188
IvID-UNet	0.5198±0.0483	0.4963±0.0590	0.7098±0.0501	0.3772±0.0097	0.4054±0.0095	0.4073±0.0092	0.4860±0.0210
IT-UNet w/o IDB	0.4875±0.0388	0.4750±0.0550	0.6617±0.0416	0.3614±0.0185	0.3931±0.0090	0.3902±0.0149	0.4615±0.0223
IT-UNet w/o ITM	0.5049±0.0313	0.4816±0.0423	0.6630±0.0260	0.3712±0.0118	0.3986±0.0061	0.3960±0.0205	0.4692±0.0149
IT-UNet (Ours)	**0.4830±0.0402**	**0.4603±0.0522**	**0.6590±0.0305**	**0.3471±0.0143**	**0.3688±0.0093**	**0.3779±0.0154**	**0.4494±0.0155**

The bold values represent the best results.

TABLE VI
QUANTITATIVE RESULTS OF ABLATION STUDY ON THE SCMC DDH DATASET WITH SDR (UNIT: %)

Algorithm	SDR (%) ↑		
	0.5 mm	1.0 mm	1.5 mm
U-Net	67.12±1.62	92.48±0.83	97.14±0.71
CvT-UNet	68.74±1.97	92.76±1.06	97.12±0.71
CvD-UNet	68.12±1.68	92.40±1.01	96.91±0.76
IvID-UNet	68.38±1.72	92.90±0.34	96.93±0.89
IT-UNet w/o IDB	70.38±2.10	92.98±0.89	97.29±0.79
IT-UNet w/o ITM	69.45±1.01	93.02±0.56	97.17±0.59
IT-UNet (Ours)	**71.19±1.76**	**93.45±1.07**	**97.31±0.56**

The bold values represent the best results.

of our hip landmark detection model, reaffirming its robust interpretability.

VI. DISCUSSION

In this work, we propose an IT-UNet model to detect six critical anatomical landmarks in hip ultrasound images for subsequent DDH diagnosis. The experimental results on two DDH datasets demonstrate the effectiveness of the proposed IT-UNet.

It is well known that ultrasound images are prone to various factors of variability during image acquisition, such as operator experience, type of device and transducer, probe orientation, different parameters, patient condition, which make the images significant different [47]. Therefore, it is important for the ultrasound-based CAD to have good generalization. To this end, we evaluate the performance of the proposed IT-UNet on two DDH datasets from different hospitals. As shown in Fig. 5, there ultrasound images have obvious visual difference. Although the diversity of ultrasound images increases the difficulties to accurately detect six landmarks, the proposed IT-UNet consistently outperforms all the comparison algorithms on both datasets. Moreover, the results of generalization study in Table VII indicate that the IT-UNet has superior generalization to other algorithms, mainly due to the proposed ITM and IDB.

In this work, a novel ITM is proposed to combine the long-range modeling capability of Transformer and the positional awareness of involution. The ITM specially generates the query, key, and value vectors by the involution projection, which can incorporate the spatial information into each token. Meanwhile, the FFN is further improved by involution layers with the SE-Network [38], which aims to fuse the hierarchical spatial features and channel-wise information. By introducing the ITM into the U-Net network, the model can well learn both the spatial-related and long-range feature representations to further improve the detection performance. Specifically, the proposed IT-UNet outperforms the previous hip landmark detection algorithm for DDH with ultrasound images in [8], which is also an encoder-decoder architecture with a specially designed dependency mining module for capturing long-range information within hip images. We think that the proposed ITM can capture more long-range information than the dependency mining module in [8]. Moreover, compared to other U-Net-based algorithms, including the original U-Net, UNet++, TransUNet, DA-TransUNet, and SCUNet++, the proposed IT-UNet also achieves the best detection performance, mainly because it integrates the ITM into U-Net for enhancing feature representations.

Although the embedded ITM in IT-UNet can effectively capture the long-range and spatial information in hip ultrasound images, it inevitably increases the computational complexity. The higher values of Params and FLOPs then lengthen the training time. However, the proposed IT-UNet requires only about 0.06s to predict six landmarks from one hip ultrasound image during the testing stage. Moreover, it also achieves the best detection accuracy compared to other algorithms. Therefore, the proposed IT-UNet has a superior trade-off between the accuracy of hip landmark detection and running time. Moreover, the model can be further optimized for faster runtime through deployment strategies, such as model pruning and quantization, so as to be more suitable for the real-time applications in clinical practice.

On the other hand, the encoder-decoder architecture generally suffers from the issue of fine-grained information loss during downsampling process. Existing methods demonstrate the effectiveness of replacing traditional maxpooing with convolutions [18]. Thus, considering the importance of positional information of landmarks in hip ultrasound images, a new IDB is developed by combining the involutions and convolutions. The Involutions in the IDB are adopted to reduce the dimensionality of input feature maps, which can preserve the valuable and detailed information. The results in Tables V and VI indicate the effectiveness of IDB. Moreover, both the ITM and IDB significantly contribute the improvement of the proposed IT-UNet on the landmark detection task.

TABLE VII
QUANTITATIVE RESULTS OF GENERALIZATION STUDY ON MRE AND SDR

Algorithm	MRE (mm) ↓							SDR (%) ↑		
	L_1(APP)	L_2(IEP)	L_3(LIP)	L_4(BRIP)	L_5(ARJP)	L_6(GLP)	Avg	0.5 mm	1.0 mm	1.5 mm
U-Net [12]	0.7001	0.5895	1.4040	0.4450	0.4809	0.4707	0.6817	53.00	80.67	91.08
DM-ResNet [8]	0.8662	0.7501	1.6099	0.5076	0.5439	0.6244	0.8170	39.00	73.25	88.17
HRNet [43]	0.7104	0.6125	1.3394	0.4548	0.5029	0.5641	0.6974	48.42	82.00	91.08
UNet++ [13]	0.7062	0.6080	**1.3153**	0.4537	0.4903	0.5224	0.6826	49.00	81.83	91.22
TransUNet [14]	0.8007	0.6815	1.4505	0.4674	0.5184	0.5221	0.7401	45.17	79.75	90.25
FARNet [16]	0.7330	0.6351	1.3379	0.4507	0.4932	0.5210	0.6951	48.67	80.75	90.92
DA-TransUNet [44]	0.8040	0.5978	1.3154	0.4431	0.4863	0.5049	0.6919	49.17	81.25	90.90
SCUNet++ [45]	0.8157	0.6922	1.5869	0.4812	0.5066	0.5306	0.7689	43.50	76.33	88.33
IT-UNet (Ours)	**0.6531**	**0.5516**	1.3170	**0.4316**	**0.4802**	**0.4539**	**0.6479**	**55.92**	**84.25**	**91.58**

The bold values represent the best results.

TABLE VIII
PARAMS AND FLOPS OF DIFFERENT ALGORITHMS FOR HIP LANDMARK DETECTION

Model	Params (M)	FLOPs (G)
U-Net [12]	6.823	48.707
DM-ResNet [8]	46.790	99.810
HRNet [43]	63.558	334.136
UNet++ [13]	9.163	34.914
TransUNet [14]	66.831	33.710
FARNet [16]	61.191	18.334
DA-TransUNet [44]	94.511	33.287
SCUNet++ [45]	43.541	16.858
IT-UNet (Ours)	181.999	94.411

In fact, the spatial-specific characteristic of involution operation makes it have significant value in exploring the spatial information in neural networks. Existing works have demonstrated its effectiveness by introducing the involution operation into the CNN, multilayer perception, and attention models for various vision tasks [24], [25], [26], [27], [28]. We believe that the involution has the feasibility to be integrated into other effective networks, such as graph neural networks (GCN), diffusion model, and recently proposed Mamba model [48], so as to further improve their feature representations. In future works, we will study the Involution-based GCN to explore more spatial relations among landmarks to further improve the landmark detection accuracy for hip ultrasound images.

Furthermore, the proposed IT-UNet also has promising feasibility for other landmark detection tasks in different medical imaging modalities, such as hip X-ray landmark detection [49], cephalometric landmark detection [50], and spine posterior corner detection [51]. In fact, we think that the proposed IT-UNet has the potential to replace the U-Net based landmark detection models, or the developed ITM and IDB can be integrated into other U-Net based models to further improve the detection performance. In future work, we will apply the IT-UNet to more landmark detection tasks to extend its application.

Despite the effectiveness of IT-UNet in this work, it still has room for improvement. The IT-UNet only detects six landmarks in this work, and its performance to detect more landmarks should be further evaluated. For example, the nuclei detection task in histopathological whole slide images with huge sizes is very difficult, we should improve the efficiency and effectiveness of IT-UNet on such a complex task. Moreover, since the IT-UNet is developed for landmark detection in 2D ultrasound images, it currently cannot be directly applied to 3D medical images. Thus, we will improve the IT-UNet for more landmark detection tasks in different imaging modalities in future work.

VII. CONCLUSION

In conclusion, we propose a novel Involution Transformer based U-Net network (IT-UNet) to promote the performance of landmark detection in the hip ultrasound images. Particularly, an Involution Transformer module is developed to capture both spatial-related information and long-range dependencies around hip landmarks. Meanwhile, the Involution Downsampling block is specifically designed to reduce the loss of valuable information in ultrasound images. The experimental results demonstrate the effectiveness of the proposed IT-UNet on two real-world datasets of infantile DDH, indicating its potentially clinical application.

REFERENCES

[1] S. Sioutis et al., “Developmental dysplasia of the hip: A review,” *J. Long-Term Effects Med. Implants*, vol. 32, no. 3, pp. 39–56, 2022.

[2] S. K. Storer and D. L. Skaggs, “Developmental dysplasia of the hip,” *Amer. Fam. Physician*, vol. 74, pp. 1310–1316, 2006.

[3] A. Kitay et al., “Ultrasound is an alternative to X-ray for diagnosing developmental dysplasia of the hips in 6-month-old children,” *HSS J.*, vol. 15, no. 2, pp. 153–158, 2019.

[4] R. Graf, “Fundamentals of sonographic diagnosis of infant hip dysplasia,” *J. Pediatr. Orthop.*, vol. 4, no. 6, pp. 735–740, 1984.

[5] D. Golan et al., “Fully automating Graf’s method for DDH diagnosis using deep convolutional neural networks,” in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2016, pp. 130–141.

[6] S. W. Lee et al., “Accuracy of new deep learning model-based segmentation and key-point multi-detection method for ultrasonographic developmental dysplasia of the hip (DDH) screening,” *Diagnostics*, vol. 11, no. 7, 2021, Art. no. 1174.

[7] X. Hu et al., “Joint landmark and structure learning for automatic evaluation of developmental dysplasia of the hip,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 345–358, Jan. 2022.

[8] J. Xu et al., “Hip landmark detection with dependency mining in ultrasound image,” *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3762–3774, Dec. 2021.

[9] J. Liu et al., “Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network,” *Biomed. Signal Process.*, vol. 86, 2023, Art. no. 105150.

[10] M. Juneja et al., “A review on cephalometric landmark detection techniques,” *Biomed. Signal Process. Control*, vol. 66, 2021, Art. no. 102486.

[11] S. S. Kshatri and D. Singh, “Convolutional neural network in medical image analysis: A review,” *Arch. Comput. Methods Eng.*, vol. 30, no. 4, pp. 2793–2810, 2023.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, vol. 9351, pp. 234–241.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[14] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” in *Proc. Int. Conf. Mach. Learn. Workshop Interpretable Mach. Learn. Healthcare*, 2021, pp. 1–13.

[15] Z. Li, S. Ying, J. Wang, H. He, and J. Shi, “Reconstruction of quantitative susceptibility mapping from total field maps with local field maps guided UU-net,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 2047–2058, Apr. 2023.

[16] Y. Ao and H. Wu, “Feature aggregation and refinement network for 2D anatomical landmark detection,” *J. Digit. Imag.*, vol. 36, no. 2, pp. 547–561, 2023.

[17] R. Liu et al., “An intriguing failing of convolutional neural networks and the coordconv solution,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 9605–9616.

[18] J. T. Springenberg et al., “Striving for simplicity: The all convolutional net,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[19] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.

[20] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, “Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 1, pp. 55–68, Feb. 2024.

[21] F. Shamshad et al., “Transformers in medical imaging: A survey,” *Med. Image Anal.*, vol. 88, 2023, Art. no. 102802.

[22] Q. Li, R. Zhong, X. Du, and Y. Du, “TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[23] D. Li et al., “Involution: Inverting the inherence of convolution for visual recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12316–12325.

[24] Y. Shao, J. Liu, J. Yang, and Z. Wu, “Spatial–spectral involution MLP network for hyperspectral image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9293–9310, 2022.

[25] Y. Hou et al., “Attention meets involution in visual tracking,” *J. Vis. Commun. Image Representation*, vol. 90, 2023, Art. no. 103746.

[26] S. Jain et al., “CoInNet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation,” *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3987–4000, Dec. 2023.

[27] A. A. Asiri et al., “Enhancing brain tumor diagnosis: Transitioning from convolutional neural network to involutional neural network,” *IEEE Access*, vol. 11, pp. 123080–123095, 2023.

[28] H. Xiao, L. Peng, S. Peng, and Y. Zhang, “Lung image segmentation based on Involution UNet model,” in *Proc. Int. Conf. Adv. Electron. Mater., Comput. Softw. Eng.*, 2022, pp. 184–187.

[29] A. Stamper, A. Singh, J. McCouat, and I. Voiculescu, “Infant hip screening using multi-class ultrasound scan segmentation,” in *Proc. IEEE 20th Int. Symp. Biomed. Imag.*, 2023, pp. 1–4.

[30] B. Gong et al., “Diagnosis of infantile hip dysplasia with B-mode ultrasound via two-stage meta-learning based deep exclusivity regularized machine,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 334–344, Jan. 2022.

[31] R. Gong et al., “Hybrid-supervised bidirectional transfer networks for computer-aided diagnosis,” *Comput. Biol. Med.*, vol. 165, 2023, Art. no. 107409.

[32] K. Oh, I.-S. Oh, V. N. T. Le, and D.-W. Lee, “Deep anatomical context feature learning for cephalometric landmark detection,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 3, pp. 806–817, Mar. 2021.

[33] Q. Yao et al., “Miss the point: Targeted adversarial attack on multiple landmark detection,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2020, pp. 692–702.

[34] H. Zhu et al., “You only learn once: Universal anatomical landmark detection,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 85–95.

[35] A. Stergiou, R. Poppe, and G. Kalliatakis, “Refining activation downsampling with SoftPool,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10337–10346.

[36] Y. M. Kwon et al., “Semantic segmentation by using down-sampling and subpixel convolution: DSSC-UNet,” *Comput., Mater. Continua*, vol. 75, no. 1, pp. 683–696, 2023.

[37] J. Li and W. Guan, “Patch merging refiner embedding UNet for image denoising,” *Inf. Sci.*, vol. 641, 2023, Art. no. 119123.

[38] J. Hu, L. Shen, and G. Sun, “ Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[39] H. Wu et al., “Cvt: Introducing convolutions to vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[40] I. A. Usmani et al., “Cartesian product based transfer learning implementation for brain tumor classification,” *Comput., Mater. Continua*, vol. 73, pp. 4369–4392, 2022.

[41] D. J. Zhang et al., “Morphmlp: A self-attention free, MLP-like backbone for image and video,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, vol. 13695, pp. 230–248.

[42] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[43] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[44] G. Sun et al., “DA-TransUNet: Integrating spatial and channel dual attention with transformer U-Net for medical image segmentation,” *Front. Bioeng. Biotechnol.*, vol. 12, 2024.

[45] Y. Chen et al., “SCUNet++: Swin-UNet and CNN bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism CT image segmentation,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7759–7767.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[47] L. Duron et al., “Can we use radiomics in ultrasound imaging? Impact of preprocessing on feature repeatability,” *Diagn. Interventional Imag.*, vol. 102, no. 11, pp. 659–667, 2021.

[48] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” 2023, *arXiv:2312.00752*.

[49] C. Liu, H. Xie, S. Zhang, Z. Mao, J. Sun, and Y. Zhang, “Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip,” *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3944–3954, Dec. 2020.

[50] H. Zhu et al., “UOD: Universal one-shot detection of anatomical landmarks,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 24–34.

[51] J. Yi, P. Wu, Q. Huang, H. Qu, and D. N. Metaxas, “Vertebra-focused landmark detection for scoliosis assessment,” in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 736–740.