NHBS-Net: A Feature Fusion Attention Network for Ultrasound Neonatal Hip Bone Segmentation

Ruhan Liu, Mengyao Liu, Bin Sheng, *Member, IEEE*, Huating Li, Ping Li, *Member, IEEE*, Haitao Song, Ping Zhang, *Senior Member, IEEE*, Lixin Jiang, and Dinggang Shen, *Fellow, IEEE*

***Abstract*— Ultrasound is a widely used technology for diagnosing developmental dysplasia of the hip (DDH) because it does not use radiation. Due to its low cost and convenience, 2-D ultrasound is still the most common examination in DDH diagnosis. In clinical usage, the complexity of both ultrasound image standardization and measurement leads to a high error rate for sonographers. The automatic segmentation results of key structures in the hip joint can be used to develop a standard plane detection method that helps sonographers decrease the error rate. However, current automatic segmentation methods still face challenges in robustness and accuracy. Thus, we propose a neonatal hip bone segmentation network (NHBS-Net) for the first time for the segmentation of seven key structures. We design three improvements, an enhanced dual attention module, a two-class feature fusion module, and a coordinate convolution output head, to help segment different structures. Compared with current state-of-the-art networks, NHBS-Net gains outstanding performance accuracy and generalizability, as shown in the experiments. Additionally, image standardization is a common need in ultrasonography. The ability of segmentation-based standard plane detection is tested on a 50-image standard dataset. The experiments show that our method can help healthcare workers decrease their error rate from 6%-10% to 2%. In addition, the segmentation performance in another ultrasound dataset (fetal heart) demonstrates the ability of our network.**

***Index Terms*— Neonatal hip bone segmentation, self-attention mechanism, medical image segmentation.**

Manuscript received October 19, 2020; revised March 30, 2021 and May 27, 2021; accepted June 06, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61872241, Grant 61572316, and Grant 81771850, in part by the Science and Technology Commission of Shanghai Municipality under Grant 18410750700 and Grant 17411952600, in part by the SJTU Medicine Engineering Interdisciplinary Research Fund under Grant YG2017MS19, and in part by The Hong Kong Polytechnic University under Grant P0030419, Grant P0030929, and Grant P0035358. *(Ruhan Liu and Mengyao Liu contributed equally to this work.) (Corresponding authors: Bin Sheng, Lixin Jiang, and Dinggang Shen.)*

R. Liu and B. Sheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

M. Liu and L. Jiang are with the Department of Ultrasound, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China; and also with the Shanghai Institute of Ultrasound in Medicine, Shanghai 200233, China (e-mail: jinger_28@sina.com).

H. Li is with the Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China (e-mail: huarting99@sjtu.edu.cn).

P. Li is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: p.li@polyu.edu.hk).

H. Song is with the Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: allen5@sjtu.edu.cn).

P. Zhang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA; and also with the Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio 43210, USA (e-mail: zhang.10631@osu.edu).

D. Shen is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China; with the Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China; and also with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: dinggang.shen@gmail.com).

I. Introduction

DEVELOPMENTAL dysplasia of the hip (DDH) is a common disease that is often found in infants. The prevalence rate of DDH is as high as 3% [1]. Identifying DDH in the early stage is essential because treatments such as the Pavlik harness are available. DDH that is not diagnosed in time can influence the quality of the patient's whole life. Moreover, serious consequences can occur, including secondary anatomical changes and leg length discrepancies, sometimes even necessitating replacement of the entire hip joint. Some researchers claim that DDH accounts for 30% of hip replacements in patients under 60 years old [2]. Thus, early and timely diagnosis of DDH is essential to maintain the quality of life of these patients.

Due to its convenience and the fact that it does not use radiation, 2-D ultrasound has become a standard test for the early diagnosis of DDH. However, 2-D ultrasound scans of the neonatal hip are limited by their technical difficulty; for example, proficiency in using Graf method [3], which are a regular and popular measurement for the clinical diagnosis of DDH, requires much professional training and technical guidance. Moreover, the examination of DDH is often completed by healthcare workers without enough experience to master standardization and measurement techniques. In [4], the researchers mentioned that incorrect anatomical identification and invisible landmarks are the main cause of incorrect diagnosis.

The ultrasound image used in diagnosis requires standard detection of key structures, including anatomical identification and usability check (Fig. 1 A). For anatomical identification, all structures mentioned should be visible and identified (checklist 1). In the usability check, three conditions need to be met (checklist 2). Fig. 1 B shows some incorrect diagnoses (a false positive sample and a false negative sample) caused by substandard images. An automatic standard detection method based on segmentation results could help healthcare workers

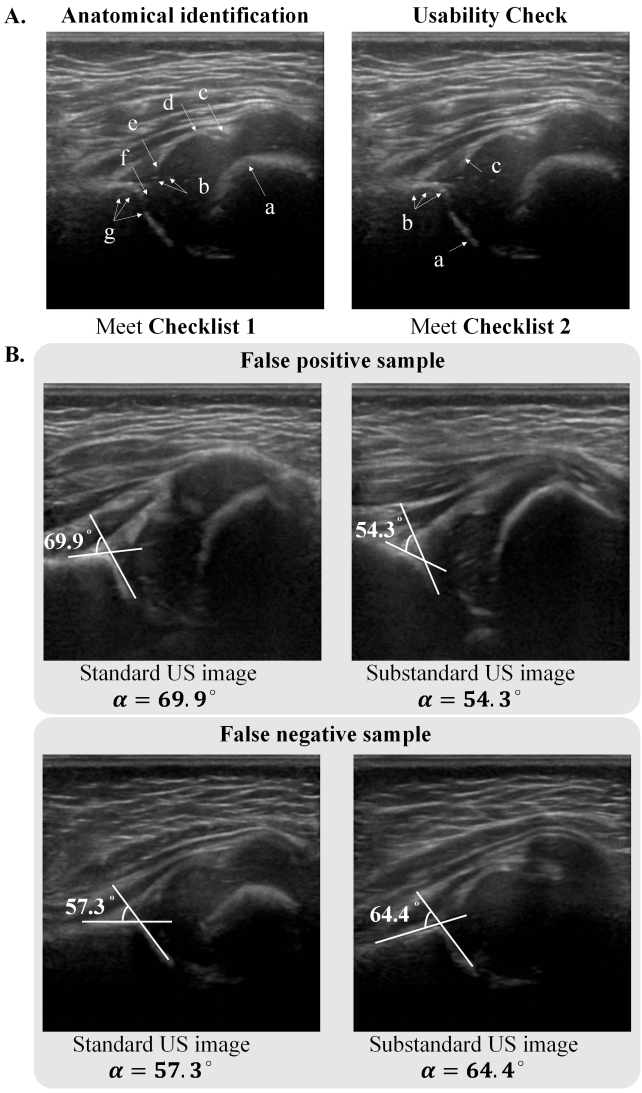

Fig. 1. DDH measurement of standard plane detection criteria and errors caused by the substandard plane. **A.** Anatomical identification (checklist 1: a) chondro-osseous border (CB), b) femoral head (FH), c) synovial fold (SF), d) joint capsule & perichondrium (JCP), e) labrum (La), f) cartilagineous roof (CR), g) bony roof (BR) and bony rim (concavity-convexity)) and usability check (checklist 2: a) the lower limb of the os ilium is visible b) the middle of the bony roof (middle plane) is parallel; c) the labrum is visible) of standard planar detection. **B.** False DDH measurements caused by substandard ultrasound images (false positive cases and false negative cases).

perform anatomical identification and usability check. The accurate segmentation of key structures provides a foundation for automatic standard plane detection. Meanwhile, since manual segmentation of key structures is laborious and time-consuming, accurate, robust, and effective automatic segmentation of key structures is necessary.

Methods that are currently used in ultrasound image-based neonatal hip joint structure segmentation are introduced. Specific feature extraction methods such as confidence-weighted structured phase symmetry (CSPS) and shadow peak (SP) features are used in this segmentation process [5]–[7]. Although these methods do provide a result, they still face challenges in terms of robustness and generalizability to new

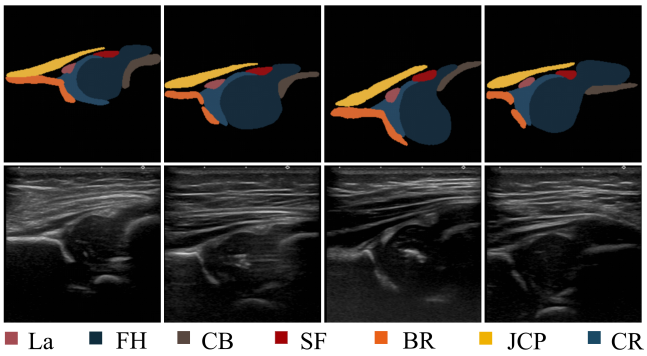

Fig. 2. Ultrasound images in the neonatal hip joint and relative manual pixel-level annotations of seven structures, including CB (dark brown), FH (dark blue), CR (blue), SF (dark red), La (light coral), BR (orange), and JCP (yellow).

data. Furthermore, they rely on manually extracted features and can only distinguish the labrum and the surface of the bony roof. These methods make it even more challenging to segment all seven key structures because of the difference in structures.

Many novel learning-based methods, such as convolutional neural networks (CNNs), obtain outstanding performance [8]–[12]. In medical image analysis, CNNs also achieve excellent results, such as in breast tumor detection, fetal ultrasound segmentation, and prostate segmentation [13]–[16]. In the segmentation of hip joint structures, some deep learning attempts are also being used. The self-attention mechanism is famous for increasing the network ability by focusing on essential features and suppressing unnecessary features. Many CNNs in medical image segmentation benefit from using attention modules [17]–[20]. Furthermore, CNNs can implicitly learn to encode absolute position information. In [21], a comprehensive set of experiments showed the validity of this hypothesis. Although current CNNs have proven to be able to learn a certain degree of location information implicitly, explorations of the use of absolute position information to improve model performance are needed. CoordConv [22] is a method that attempts to use absolute position information as much as possible and has achieved outstanding performance. In medical imaging analysis, there are also some methods that use CoordConv to improve performance [23].

In our task, an attention-based segmentation network backbone can be considered for the construction of pediatric hip segmentation-specific networks due to the advantages mentioned above. In particular, the dual attention model based on [24], due to its superior correlation position enhancement capability, can be used as the backbone. Furthermore, since DDH diagnosis has a high requirement for the recognition of structural edges, the fusion model based on [25] is also considered for integrating different feature maps to improve the accuracy of the edges. Moreover, as shown in Fig. 2, the seven structures of the hip joint show a position correlation. Using CoordConv [22] in the output head to encode absolute position information is a potential way to improve the ability of the segmentation network.

In our work, we make the following contributions:

1) We propose NHBS-Net, the first framework to segment the seven key structures of the neonatal hip joint. The results illustrate that compared with other segmentation networks, NHBS-Net can effectively improve the segmentation performance of the Dice similarity coefficient (DSC) and Hausdorff distance (HD).
2) We design an enhanced dual attention module (EDAM) using the location attention module (LAM) and enhanced channel attention (ECA), which develops an enhanced channel to learn the weights of different channels. EDAM can learn global feature correlations and importantly improve performance.
3) To improve the accuracy of the structure edges, we develop a two-class feature fusion module (2-class FFM) containing two fusion parts. First-class feature fusion module (first-class FFM) can effectively merge the location and channel attention maps, and second-class feature fusion module (second-class FFM) can fuse the low-level features and high-level attention features.
4) We introduce a location-related output head, the coordinate convolution output head (CCOH), to generate segmentation results from the extracted features. The absolute position information can be encoded into the feature map to reduce the error segmentation caused by structural similarity.

II. Related Work

A. Neonatal Hip Bone Structure Segmentation

Many studies have developed methods in ultrasound neonatal hip joint segmentation to automatically distinguish different key structures, such as ilium and acetabulum bone surfaces. Several manual features have been introduced for segmenting acetabulum bone surfaces. Quader et al. [5] illustrate the confidence-weighted structured phase symmetry (CSPS) feature, which combines the near-constant acoustic properties of bone and cartilage structures. This work can reduce soft tissue false positives in hip bone segmentation. Furthermore, Pandey et al. [6] introduced the shadow peak (SP) feature. This simplified feature uses only bone shadowing features to segment bone. Hasan et al. [7] recently demonstrated a framework that uses particle swarm optimization (PSO) and the statistical level set (SLS) method to segment ilium and acetabulum bone surfaces. PSO is used to determine the locations of the initial contour and the region of interest (ROI), and the SLS method is developed for segmenting the essential anatomical structures from the ROI. The disadvantages of these models are that they can only segment ilium and acetabulum bone surfaces with similar features, and the methods have only been tested on a small dataset. Thus, highly manually crafted features remain challenges for the robustness and generalizability of this method to large datasets and new data.

Deep learning methods are proposed for this task due to their outstanding performance in medical image analysis. In [26], a method using superpixel classification with a CNN achieved an HD of 2.1 ± 0.9 mm between contours. In [27], a Mask R-CNN [28]-based framework was introduced in acetabulum bone surface segmentation, and the net gain was 0.386 in the DSC metric. [27] also compared their model with U-Net [29] and fully convolutional neural network (FCN) [11]. The DSCs of U-Net and FCN in their dataset were 0.049 and 0.223, respectively. The DSC they obtained was very low, which may have been caused by the smaller segmentation targets and the image dataset, which was collected for a small group of patients. El-Hariri et al. [30] compared hand-engineered feature methods and deep learning methods in acetabulum bone surface segmentation. In their results, the grayscale input U-Net gained 0.86 and 0.92 in DSC for the two datasets they used. However, these studies focused only on the segmentation of ilium and acetabulum bone surfaces. Segmenting the seven key structures has not yet been discussed due to the structural similarity and complexity.

B. Semantic Segmentation and Self-Attention Modules

Pixel-level image segmentation has recently become fast and precise because of the development of deep learning methods. In 2014, a FCN was proposed [11] that used a deconvolutional layer to replace the fully connected layer. Based on the FCN structure, encoder-decoder frameworks such as Seg-Net [31] and U-Net [29] are used. The shortcut connection between the encoder layer and the relative decoder layer helps in the reconstruction of details. In DeepLab networks [12], [32], atrous convolution and conditional random fields are used to improve the segmentation performance. The atrous spatial pyramid pooling (ASPP) module based on atrous convolution replaces basic pooling and can maintain spatial resolution. Moreover, the latest version of DeepLab (V3 plus) [32] focuses more on a decoder that uses the low-level features and high-level features to recover segmentation labels very well. Additionally, Mask R-CNN [28] is a two-stage instance segmentation method, and it can be regarded as a combination of object detection and semantic segmentation. In Mask R-CNN, a feature pyramid network [33] is used to extract multiscale features.

The self-attention mechanism has achieved substantial success in NLP [34], and recently, many studies [17], [24], [25] have made attempts to use it in segmentation tasks in computer vision. Many attention-based models have achieved excellent results in medical image segmentation [17], [20]. In [17], the author proposed a novel attention-guided dense-upsampling network with an asymmetrical encoder-decoder structure and introduced an attention-guided dense-upsampling block. In [20], a 3-D self-attention network that can capture a wide range of spatial information was mentioned. Due to the current use of 2-D ultrasound for pediatric hip segmentation, the existing methods face challenges in universality and generalization. Furthermore, the performance of deep learning methods is affected by insufficient data.

III. Method

The architecture of our NHBS-Net has four parts: a feature extraction module based on dilated ResNet [35], [36], EDAM to generate location attention maps and enhanced channel attention maps, 2-class FFM to fuse location attention and channel attention features in first-class FFM and integrate

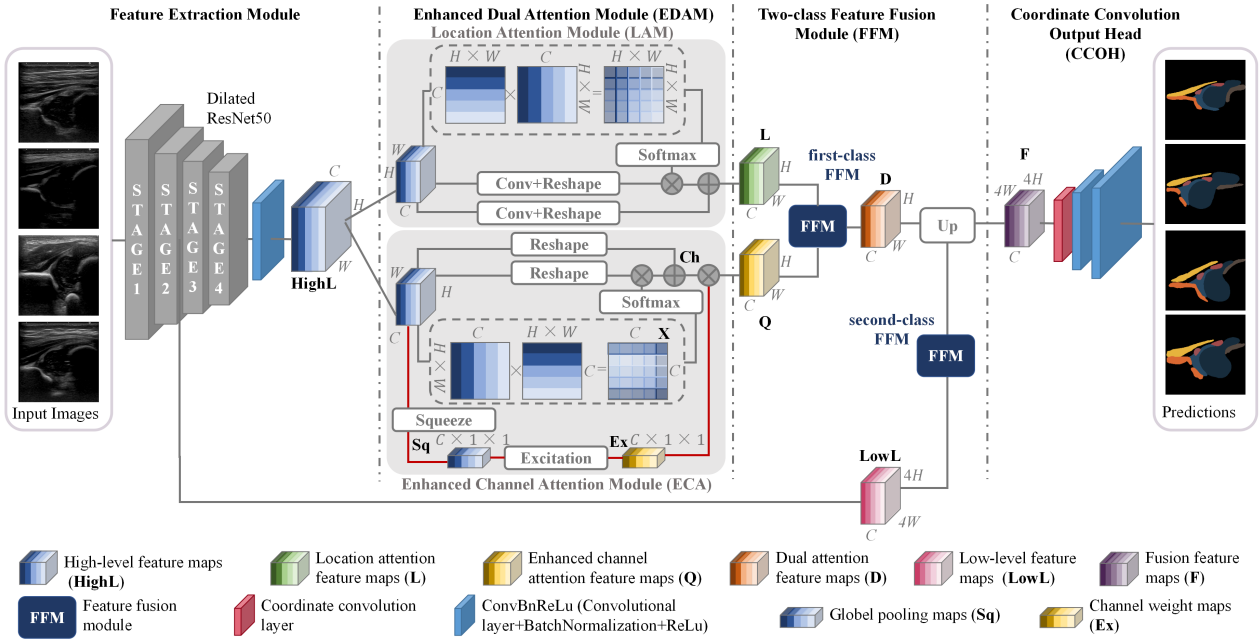

Fig. 3. Overview of the NHBS-Net. The ultrasound images are passed through a feature extraction module to obtain high-level features ($\boldsymbol{HighL}$) and low-level features ($\boldsymbol{LowL}$). EDAM is used on the high-level features, which generates two-path enhancement features named location attention features ($\boldsymbol{L}$) and enhanced channel attention features ($\boldsymbol{Q}$). The two kinds of attention features use the first-class feature fusion module to obtain dual attention features ($\boldsymbol{D}$), which generate the fusion features ($\boldsymbol{F}$) with low-level features through the second-class fusion module. The CCOH generates the final segmentation predictions based on the fusion features.

high-level and low-level feature maps in second-class FFM, and CCOH to decode feature maps and generate predictions. The structure of the NHBS-Net model is shown in Fig. 3. EDAM can learn the global feature correlations to reduce the segmentation errors of different structures. The 2-class FFM integrates low-level and high-level features, making the segmentation results at the edge superior. Our CCOH can reduce the error segmentation of similar features in different structures.

A. Feature Extraction Module and Enhanced Dual Attention Module (EDAM)

The feature extraction model is based on a dilated ResNet architecture. The dilated residual network (DRN) [36] is a famous CNN structure that uses dilated convolution to reduce the loss of spatial information caused by downsampling in feature extraction. In our network, the feature extraction module's output is the set of high-level feature maps $HighL$, and the output of the 2^{nd} stage in dilated ResNet forms the low-level feature maps $LowL$.

Understanding global feature correlations in the feature maps is helpful for distinguishing structures. Two-path attention modules [24], [25], [37]–[39] are widely used in segmentation tasks and obtain outstanding performance. Dual attention (DA) [24] can capture long-range contextual information dependencies to help study the information of different key structures, especially location information dependency. After the feature extraction model, an EDAM revised by DA [24] is adopted to enhance the network's ability to understand context information. In the location attention part of our NHBS-Net, we adopt the same structure as [24]. After passing through the location path, the location attention maps L are generated.

Furthermore, we propose enhanced channel attention (ECA) based on the channel attention module introduced in [24]. An enhanced channel introduced in [37] was added to learn the weights of different channels of the feature maps (the red line in Fig. 3.). The high-level feature maps $HighL \in \mathbb{R}^{C\times H\times W}$ are directly reshaped into $HighL \in \mathbb{R}^{C\times N}$ and $HighL \in \mathbb{R}^{N\times C}$ (N is $H \times W$). After that, matrix multiplication is performed in $HighL \in \mathbb{R}^{C\times N}$ and $HighL \in \mathbb{R}^{N\times C}$. Next, the multiplication result is fed into a softmax layer to calculate the channel attention map $X \in \mathbb{R}^{C\times C}$:

$$X_{ji} = \frac{exp(HighL_i \cdot HighL_j)}{\sum_{i=1}^{C} exp(HighL_i \cdot HighL_j)} \tag{1}$$

where X_{ji} represents the effect of the i^{th} channel on the j^{th} channel. Moreover, we perform another matrix multiplication between the $HighL \in \mathbb{R}^{N\times C}$ and $X \in \mathbb{R}^{C\times C}$. The multiplication output of X and $HighL$ is $Ch \in \mathbb{R}^{C\times H\times W}$, which can be calculated by:

$$Ch_j = \sum_{i=1}^{C}(HighL_{ji}X_i) \tag{2}$$

In the enhanced channel, $HighL \in \mathbb{R}^{C\times H\times W}$ is fed into a global pooling layer to obtain squeeze attention $Sq \in \mathbb{R}^{1\times 1\times C}$. Sq is:

$$Sq = \frac{1}{N}\sum_{s=1}^{H}\sum_{t=1}^{W} HighL_{st} \tag{3}$$

Then, the squeeze attention Sq is inputted into the excitation operator, which uses two-layer convolution with ReLU and a sigmoid activation function respectively to gain excitation attention $Ex \in \mathbb{R}^{1\times1\times C}$. Ex is:

$$Ex = \sigma(W_2 ReLu(W_1 Sq)) \tag{4}$$

where W_1 and W_2 are the weights of the two convolution layers.

Then, the reshaped multiplication result can be calculated by Ch and Ex:

$$Q_j = \beta \cdot (\sum_{i=1}^{C} Ch_{ji} \cdot Ex_i) + HighL_j \tag{5}$$

where β is the weight of the learning path and the channel attention maps Q are weighted sums of the features across all channel and high-level feature maps $HighL$.

B. Two-Class Feature Fusion Module (2-Class FFM)

In the 2-class FFM, we design to use two FFM structures [25] to fuse different paths of attention maps and different levels of feature maps. The first-class FFM is used to fuse location path attention maps and enhanced channel path attention maps. The first-class FFM replaces the sum fusion module in [24], which is suitable for dual attention path fusion. Furthermore, the second-class FFM is applied to integrate high-level and low-level features. In [25], the same FFM was used to fuse context paths and spatial paths. The features in the spatial path represent low-level features, which a different shallow CNN extracts. In our work, the low-level features fused in the second-class FFM are the output of the 2^{nd} stage of the dilated ResNet backbone, which means that we do not need additional shallow CNNs for low-level feature extraction. The specific implementation details of the 2-class FFM are given below.

Due to the difference of the two path attention feature maps obtained by EDAM, directly concatenating them may not maintain important information properly. We designed a first-class FFM used after EDAM to integrate location attention maps L and enhanced channel attention maps Q to form dual attention maps D. The dual attention maps D are calculated by:

$$D_j = \sum_{i=1}^{C} DK_{ji} \cdot DM_i + DK_j \tag{6}$$

In Eq. (6), $DK = ReLu(W_3(L \oplus Q))$, $DM = \sigma(W_5 ReLu(W_4(\frac{1}{N}\sum_{s=1}^{H}\sum_{t=1}^{W} DK_{st})))$, and W_3, W_4, and W_5 are the weights of the convolutional layers in the first-class FFM, and $\oplus$ represents the concatenation operator.

Furthermore, it is crucial to use the extracted features' spatial information to achieve outstanding segmentation prediction. In CNNs, however, consecutive downsampling encoding high-level semantic information can lead to spatial information loss. Several research studies have applied spatial information from previous studies by extracting low-level features [12], [25]. The DeepLab (V3 plus) model [12] concatenates low-level features with high-level features, acquiring multiscale information. BiSe-Net [25] extracts low-level features by using an additional spatial path composed of convolutional layers with different kernel sizes. In our work, we develop the second-class FFM, which uses the FFM to fuse high-level dual attention maps D and low-level feature maps $LowL$ in order to generate feature fusion maps F. In this way, our network does not require additional spatial paths to extract low-level features, and NHBS-Net also integrates multiscale features through the fusion model. The feature fusion maps F are shown below.

$$F_j = \sum_{i=1}^{C} FK_{ji} \cdot FM_i + FK_j \tag{7}$$

In Eq. (7), $FK = ReLu(W_6(D \oplus LowL))$, $FM = \sigma(W_8 ReLu(W_7(\frac{1}{16N}\sum_{s=1}^{4H}\sum_{t=1}^{4W} FK_{st})))$, and W_6, W_7, and W_8 are the weights of the convolutional layers in the second-class FFM. The 2-class FFM is a combination of the first-class FFM and second-class FFM.

C. Coordinate Convolution Output Head (CCOH)

After feature extraction by the dilated ResNet module, feature enhancement by EDAM, and feature fusion by the 2-class FFM, the CCOH is used to generate the segmentation result. CNN structures, to some degree, can learn absolute location information, which is proven by the experiments in [21]. There are also some works that have attempted to use absolute position information in CNNs [22], [40]. In [22], the coordinates of each pixel in the feature maps are added as two channels. As a simple extension of the standard convolutional layer, this operation can be easily added to existing models to improve model performance. All ultrasound images have seven key structures in our task, and the relationship between the structures is associated with their locations. The absolute location information provided by the CCOH could help correct location-related segmentation errors. Thus, our module uses the coordinate convolution layer to build the decoder module to recover the segmentation labels and enhance the location information to improve performance.

D. Loss Function

For seven key structure segmentation, specific loss function needs to be used because of the different corresponding size among these structures. Therefore, in our task, the loss function is a combination loss of Focal loss [41] and cross-entropy (CE) loss. For calculating the loss, we defined the p_t as:

$$p_{kij} = \begin{cases} p, & \text{if } y_{ij} = k \\ 1-p, & \text{otherwise} \end{cases} \tag{8}$$

where $k \in \{0, 1, 2, ..., 7\}$ is the classes in segmentation, $y_{ij} \in \{0, 1, 2, ..., 7\}$ (0 represents the background and 1 to 7 represent seven key structures of hip joint) represents pixel labels of the segmentation prediction,$i \in (0, H)$, $j \in (0, W)$, H and W are the height and W is the width of the prediction labels respectively, and p is the model's assessed probability for the class with pixel label $y_{ij} = k$.

The CE loss is:

$$L_{CE} = \sum_{k=0}^{7} \sum_{i}^{H} \sum_{j}^{W} -log(p_{kij}) \tag{9}$$

Then, the focal loss is:

$$L_F = \sum_{k=0}^{7} \sum_{i}^{H} \sum_{j}^{W} -\alpha_k (1 - p_{kij})^{\gamma} log(p_{kij}) \tag{10}$$

where α_k is the class balanced weight and γ is the modulating factor. In our task parameter α_k is 0.5, and γ is 2. Thus, the combination loss function to train the model is:

$$L = L_{CE} + \beta L_F \tag{11}$$

where β is the weight of focal loss L_F.

IV. Experimental Results and Discussions

A. Datasets

1) NHU Dataset: The study was approved by the ethics committee of the Shanghai Jiao Tong University Affiliated Sixth People's Hospital. We collected 562 samples from 270 patients to form the Neonatal Hip Ultrasound (NHU) dataset with seven key structures annotated. Infants aged 0-6 months who were suspected of having DDH were included in this study. A 5/7.5 MHz linear ultrasound multifrequency transducer was employed to acquire 787 2-D ultrasound coronal images, with a 40-55 mm depth setting. The images were stored in DICOM format. All of the ultrasound images were cropped to avoid revealing sensitive patient data. Image acquisition was performed by experienced sonographers. The exclusion criteria were as follows:

1) Substandard ultrasound image slices with the following problems were excluded: incomplete/unclear display of the seven main anatomical structures and not satisfying the middle plane.
2) Images with infant hip dysplasia caused by cerebral palsy, joint contractures, purulent hip arthritis, and other diseases were excluded.
3) Images of patients with other hip joint diseases and limb deformities were excluded.
4) Ultrasound images of patients with a severity of hip dysplasia that was judged to Type III and exceeding Type III by the Graf method were excluded.

According to the exclusion criteria, 225 images were excluded. The seven anatomical landmarks of the remaining 562 images were marked and proofread by two sonographers with years of experience. We randomly split the data into a training set (204 cases containing 400 images), a validation set (20 cases containing 53 images), and a testing set (47 cases containing 110 images). In the training, validation, and test sets, 22.00%, 28.30%, and 27.27% of the ultrasound images were affected by DDH. These sets were used for the subsequent training, validation, and testing procedures and the accuracy evaluation and prediction of the DSC [42], HD [43], [44], and average Hausdorff distance (AHD).

2) SPJ Dataset: The dataset for the standard plane judgment (SPJ dataset) includes a total of 50 cases of DDH images without the seven key structures annotated, of which 25 cases are standard and 25 cases are substandard. The inclusion and scanning of the standard cases were performed in the same way as for the NHU dataset. The ratio of normal and abnormal hip joints (or types I and II) is 1:1. The substandard cases were selected from the clinical practice of the same center and were divided into the following five categories according to the possible conditions that occur in clinical practice: a) nonmiddle plane, b) lack of a chondro-osseous border, c) poor labrum display, d) incomplete bony roof or poor display of the lower limb of the os ilium, e) poor display of the joint capsule and perichondrium synovial plica; there were five images of each type. The standard plane gold standard judgment for this dataset was provided by two sonographers with more than five total years of experience in pediatric hip ultrasound. Additionally, we provide two judgments from two healthcare workers for comparison with the automatic detection method. The SPJ dataset is available by accessing: https://github.com/hidden-ops/NHBS-Net_SPJ_dataset.

3) FHU Dataset: Similar to ultrasound examination of the hip in infants, the standardization of ultrasound images is common in many ultrasound examinations. The problem of finding the standard plane usually affects the diagnostic results of the examination, and in obstetric examinations, many of the judgment standards for the ultrasound standard plane are closely related to the clear visibility of key structures [45], [46]. To evaluate the adaptability of our NHBS-Net model to ultrasound standardization issues, we collected a new dataset called the Fetal Heart Ultrasound (FHU) dataset. The FHU dataset is composed of 128 four-chamber view standard planes of the fetal heart from 50 pregnant women with a fetal gestational age of 16-25 weeks. The fraction of patient ultrasound images that show fetal heart abnormality is 28.1%. All subjects received ultrasonic examinations between January 2018 and December 2020 at the Shanghai Jiao Tong University Affiliated Sixth People's Hospital. The ultrasound images were acquired by an experienced sonographer with a low-frequency convex array probe at 3.5 MHz. In the FHU dataset, 128 samples from 50 patients had five key structures annotated, including the left atrium, right atrium, left ventricle, right ventricle, and aorta. To test the model performance, we divided the FHU dataset into the FHU training dataset (including 76 images of 26 infants), the FHU validation dataset (including 28 images of 11 infants), and the FHU test dataset (including 28 images of 13 infants). The training, validation, and testing processes were the same as those of the NHU dataset.

B. Implementation Details

In this segmentation task, we used the NHU dataset to train, validate, and test the models. The ultrasound images in the NHU dataset were resized to 256×256. To improve the training data diversity, we used data augmentation methods to expand the training data size. We used a gamma transform, a grayscale linear transformation, and rotations in different directions to increase the number of pictures to 9 times the original number.

TABLE I

SEGMENTATION PERFORMANCE COMPARISON BETWEEN OUR NHBS-NET AND OTHER STATE-OF-THE-ART CNNS INCLUDING SEG-NET [31], U-NET [29], FCN [11], AU-NE T [17], DA-NET [24], SCSE-NET [37], AND BISE-NET [25]. ALL OF THESE MODELS ARE TRAINED ON A 400-IMAGE NHU TRAINING SET AND TESTED ON A 110-IMAGE NHU TESTING SET.

Model name	Average Performance	Key Structures of Hip Joint						
		CB	FH	CR	SF	La	BR	JCP
Dice Similiarity Coefficient (DSC) – %								
NHBS-Net	**87.85 ± 4.66**	88.85±4.16	**92.58 ± 2.95**	**83.85 ± 5.49**	**84.50 ± 7.47**	**84.37 ± 5.97**	88.08±4.19	**92.72 ± 2.37**
Seg-Net [31]	85.44±6.24	87.31±4.93	90.56±2.91	80.67±6.41	79.74±11.73	81.82±6.76	87.45±4.53	90.53±6.42
U-Net [29]	87.22±4.90	88.96±4.10	91.97±3.14	81.55±6.10	82.93±8.48	84.27±6.34	88.23±3.81	92.62±2.33
FCN [11]	86.47±5.01	88.24±4.86	91.99±2.80	82.98±4.81	80.76±9.35	82.29±6.46	88.18±3.90	90.84±2.91
AU-Net [17]	86.80±5.05	**89.44 ± 3.75**	92.32±2.98	82.69±5.15	82.64±8.15	79.23±8.22	**88.91 ± 3.57**	92.36±3.50
DA-Net [24]	86.38±5.30	87.71±4.45	92.07±2.92	82.69±5.89	80.98±9.14	83.87±5.88	87.00±4.74	90.31±4.09
ScSE-Net [37]	87.05±5.13	88.96±4.62	91.72±3.01	83.33±5.85	81.36±8.61	83.77±5.82	88.20±4.30	92.03±3.67
BiSe-Net [25]	86.61±4.81	87.93±4.16	92.25±2.85	82.57±5.50	82.58±7.89	82.56±5.81	87.18±4.37	91.19±3.08
Hausdoff Distance (HD) – pixel								
NHBS-Net	**8.42 ± 6.30**	**7.88 ± 8.50**	12.28±7.41	16.06±13.82	**4.55 ± 3.00**	5.00±2.24	**5.32 ± 3.03**	**7.85 ± 6.11**
Seg-Net [31]	11.48±10.38	11.19±12.76	14.46±6.20	19.05±13.48	6.81±5.45	7.75±8.10	6.81±9.81	14.28±16.82
U-Net [29]	9.29±7.17	8.56±8.94	13.76±8.30	17.58±15.06	5.94±3.65	4.88±2.41	5.33±4.16	9.01±7.71
FCN [11]	9.00±6.56	8.53±8.58	13.34±7.83	15.96±12.58	4.98±3.05	**4.85 ± 1.94**	5.48±4.03	9.88±7.91
AU-Net [17]	8.98±6.33	8.56±8.34	12.68±8.17	16.88±10.46	5.72±5.19	5.65±2.46	4.86±2.88	8.47±6.85
DA-Net [24]	9.17±6.90	8.74±8.91	12.37±7.00	16.15±13.31	5.37±3.55	4.97±2.38	5.84±5.35	10.79±7.80
ScSE-Net [37]	10.86±10.96	13.29±21.71	15.75±12.68	17.19±13.88	6.26±4.29	5.32±2.42	5.84±5.47	12.40±16.27
BiSe-Net [25]	8.76±6.57	8.22±7.94	**11.82 ± 6.32**	**15.87 ± 11.04**	5.00±3.26	5.41±2.19	6.41±10.05	8.57±5.18
Average Hausdoff Distance (AHD) – pixel								
NHBS-Net	**0.32 ± 0.40**	0.31±0.54	0.28±0.27	0.69±0.96	**0.31 ± 0.59**	**0.28 ± 0.17**	**0.17 ± 0.15**	**0.14 ± 0.11**
Seg-Net [31]	0.46±0.57	0.44±0.65	0.39±0.25	0.78±0.68	0.51±0.80	0.44±0.44	0.25±0.40	0.43±0.78
U-Net [29]	0.36±0.44	0.30±0.45	0.36±0.36	0.85±1.10	0.37±0.68	0.29±0.21	0.20±0.19	0.15±0.13
FCN [11]	0.35±0.40	0.31±0.51	0.32±0.27	0.67±0.80	0.41±0.69	0.32±0.19	0.21±0.17	0.21±0.18
AU-Net [17]	0.34±0.38	**0.27 ± 0.44**	0.32±0.37	**0.61 ± 0.55**	0.39±0.72	0.42±0.27	0.18±0.13	0.17±0.20
DA-Net [24]	0.36±0.45	0.33±0.59	0.30±0.26	0.70±0.79	0.42±0.66	0.29±0.19	0.26±0.38	0.24±0.30
ScSE-Net [37]	0.43±0.72	0.64±1.76	0.37±0.34	0.76±0.94	0.42±0.68	0.30±0.18	0.23±0.38	0.31±0.78
BiSe-Net [25]	0.34±0.42	0.30±0.49	**0.28 ± 0.24**	0.62±0.56	0.36±0.58	0.33±0.20	0.30±0.78	0.17±0.12

All the experiments were implemented on an Intel XeonE5-2630 v4 @ 2.20 GHz CPU and NVIDIA GeForce RTX 2080 Ti on ArchLinux. All models were implemented in PyTorch using the root-mean-square prop algorithm with momentum. The segmentation networks were trained for 11 epochs. When validation loss (the total average loss in the validation set) did not decrease for 2 epochs, the learning rate decayed.

Furthermore, we used the grid search algorithm to determine the hyperparameters, including the batch size and initial learning rate. We explored batch sizes from 2 to 32 and initial learning rates in the range of 10^{-1} to 10^{-7}. We saved the model and regarded it as the final model when the validation set's performance (20 cases, 53 images) was the best in the 11 epochs. We tested the final model on a new test set (47 cases, 110 images) to obtain the model performance results.

In addition, we compared our NHBS-Net model with state-of-the-art segmentation algorithms, including U-Net [29], Seg-Net [31], FCN [11], AU-Net [17], DA-Net [24], BiSe-Net [25], and ScSE-Net [37]. These models adopt the same grid search method as the NHBS-Net mentioned above. The Seg-Net, FCN, and U-Net models use the same U-shape encoder-decoder structures, but FCN and U-Net have skip connections with feature maps. DA-Net uses the same feature extraction method as our model but applies an original dual attention module, a simple sum feature fusion method without the FFM blocks, and a normal output head without the CoordConv layer. BiSe-Net uses the high-level feature extracted from the ResNet [35] structure and the low-level feature obtained by another shallow CNN network. BiSe-Net uses feature fusion modules similar to those of NHBS-Net but without the dual attention module proposed for DA-Net or the CoordConv layer. AU-Net also has high-level and low-level feature pathways and a channelwise attention mechanism, similar to BiSe-Net, but introduces an attention-guided dense upsampling block. The loss function, which is a combination of focal loss and cross-entropy loss, is the same in all models.

C. Comparison with Other State-of-the-Art CNN Models on the 400-Image NHU Training Set

To compare the performance of our NHBS-Net and other state-of-the-art CNNs, we first trained the models (Seg-Net [31], U-Net [29], FCN [11], AU-Net [17], DA-Net [24], ScSE-Net [37], BiSe-Net [25], ScSE-Net [37], and our NHBS-Net) on the NHU training dataset containing 400 images of 204 cases. Then, we validated the models on the validation dataset. The best performance models were saved as the final model to represent the models' ability. Furthermore, the final models were tested on the testing data, including 110 ultrasound images from 47 cases. The metrics, including DSC, HD, and AHD, are shown in Table I. Regarding the seven structures, our NHBS-Net model gains 88.85 in CB, 92.58% in FH, 83.85% in CR, 84.50% in SF, 84.37% in La, 88.08% in BR, and 92.72% in JCP in DSC. The average DSC of all the structures is 87.85%, which is the best segmentation result among all models.

Additionally, due to the DSC's insensitivity to the boundary of structures, we also evaluated all methods on the AHD (the AHD can represent the similarity of shapes and address the problem that the HD value is easily affected by an extreme point). As shown in Table I, the average HD and AHD

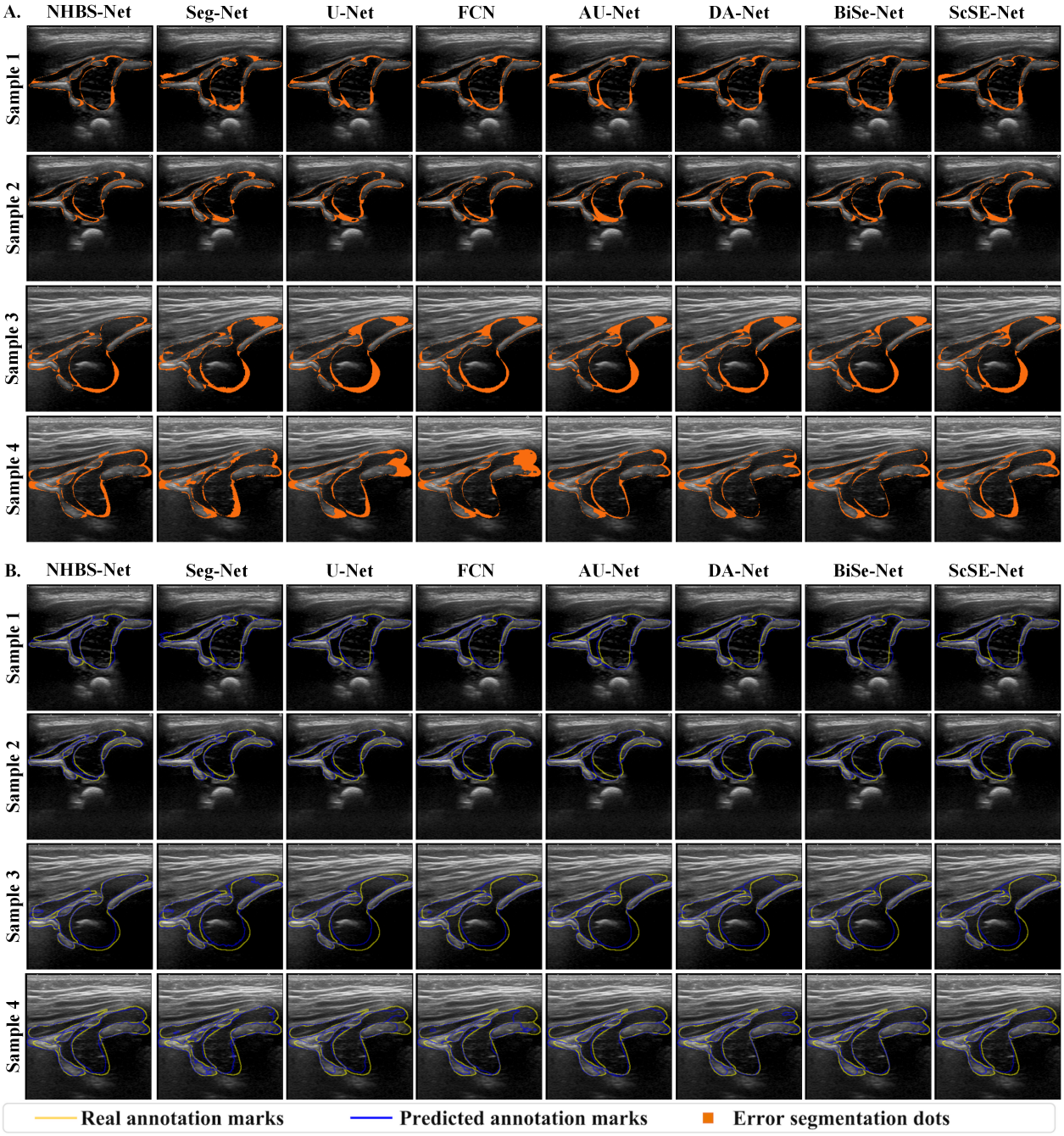

Fig. 4. Prediction diagram of Seg-Net [31], U-Net [29], FCN [11], AU-Net [17], DA-Net [24], ScSE-Net [37], BiSe-Net [25], and NHBS-Net based on four randomly selected testing set pictures. **A.** Schematic diagram of segmentation errors in different methods. **B.** Comparison chart of the actual annotations and predictions of the seven structures in different methods.

obtained by our NHBS-Net are the best among all test models. The average HD of NHBS-Net is 8.42, and the average AHD of our NHBS-Net is 0.32. Among the other novel CNN models, the best HD is 8.76, and the best AHD is 0.34. Table I shows that our NHBS-Net obtains the best results in terms of the five key structures' DSC metrics. For the HD and AHD metrics, NHBS-Net achieves the best performance on 4 structures.

Furthermore, for the DSCs of the test samples, NHBS-Net obtained 24 samples in the range of 90% to 95%, which is much higher than the other CNNs. In Seg-Net, FCN, DA-Net, U-Net, BiSe-Net, AU-Net, and ScSE-Net, the numbers are 3, 4, 7, 9, 10, 12, and 13, respectively. In the distribution of the segmentation performance of each sample, our NHBS-Net has outstanding performance. Moreover, for the average HD of all key structures, the NHBS-Net model has the lowest average HD value, which is 8.42. Seg-Net's mean HD is 11.48, U-Net's is 9.23, FCN's is 9.00, AU-Net's is 8.98, DA-Net's is 9.17, BiSe-Net's is 8.76, and ScSE-Net's is 10.86. In Fig. 4, we show the segmentation results of four randomly chosen

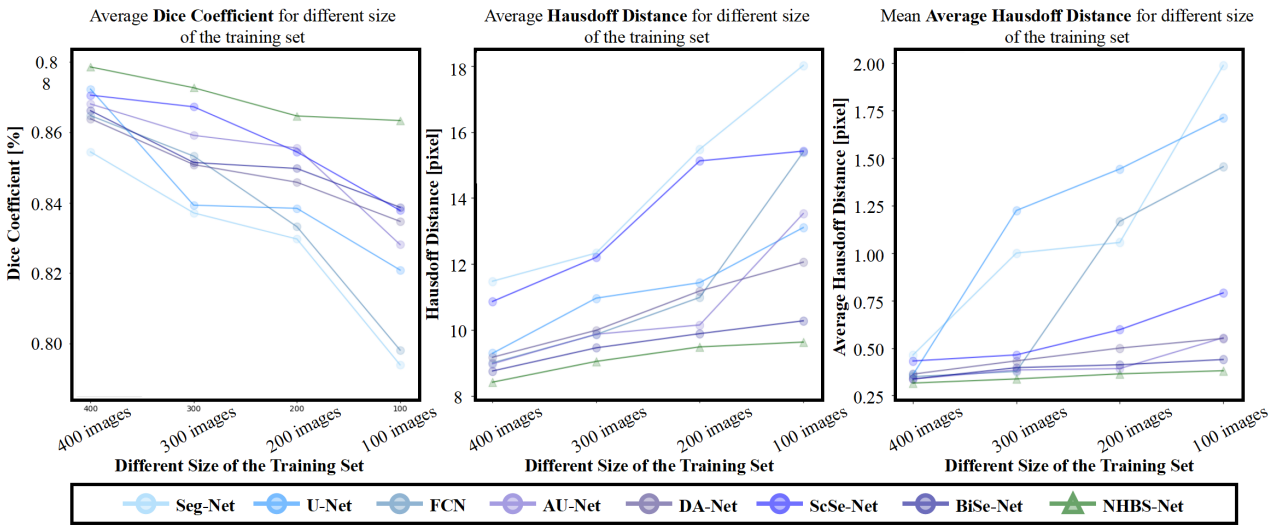

Fig. 5. The line charts show the average DSC, HD, and AHD distribution on the test set (110 images) with training size decreases of 25% (300 images), 50% (200 images), and 75% (100 images). The left, middle, and right images show the trends of the DSC, HD, and AHD, respectively, when the training set's size was reduced from 400 images to 100 images.

TABLE II
SEGMENTATION PERFORMANCE IN TERMS OF THE AVERAGE DSC (%) OF ALL KEY STRUCTURES COMPARE FOR OUR NHBS-NET AND OTHER STATE-OF-THE-ART CNNS, INCLUDING SEG-NET [31], U-NET [29], FCN [11], AU-NET [17], DA-NET [24], SCSE-NET [37], AND BISE-NET [25], UNDER A DECREASING TRAINING SIZE OF 300 IMAGES, 200 IMAGES, AND 100 IMAGES.

Model name	Pmt size (M)	Size of training set		
		300	200	100
NHBS-Net	173.4	**87.26 ± 5.00**	**86.64 ± 6.03**	**86.33 ± 5.40**
Seg-Net [31]	117.9	83.71±9.03	82.97±14.77	79.40±9.27
U-Net [29]	240.0	83.93±15.35	83.84±14.15	82.09±14.98
FCN [11]	102.7	85.32±5.07	83.32±16.37	79.82±16.19
AU-Net [17]	744.2	85.91±5.38	85.56±5.34	82.81±6.36
DA-Net [24]	198.3	85.08±5.84	84.58±5.88	83.47±6.17
ScSE-Net [37]	354.0	86.71±5.28	85.44±5.08	83.78±6.66
BiSe-Net [25]	123.2	85.14±5.16	84.97±5.26	83.87±5.93

samples. Both the error areas and contour lines show the superiority of our NHBS-Net.

D. Comparison with Other CNN Models for Decreasing Training Data Size

The size of the training dataset is closely linked with the performance of deep learning models. However, enlarging the dataset will bring cumbersome labeling work. Labeling in our task is time consuming and laborious to achieve pixel-level labeling of the seven structures. Our NHBS-Net proposes an ECAM that can learn the global feature correlation to reduce the segmentation errors of different structures. The ECA in the ECAM combines an SE-block [38] with the channel attention module proposed in [24], focusing attention on meaningful feature maps. The 2-class FFM integrates low-level and high-level features, preserving more detailed information. The CCOH using the absolute position information can reduce the erroneous segmentation of similar features in different structures. Therefore, we conduct experiments to explore the impact of data reduction on the model performance, validating the generalizability of our model. We randomly cut 25%, 50% and 75% of the samples from the training set to build 25%, 50%, and 75% training sets, used these training data to train the models (NHBS-Net, Seg-Net, U-Net, FCN, AU-Net, DA-Net, ScSE-Net, BiSe-Net, and ScSE-Net), and evaluated these models on the testing set of 110 unchanged images.

In Table II, the DSCs of NHBS-Net for all sizes of training are the best. NHBS-Net are 0.55%, 1.08%, and 2.46% higher than the best-performing CNN models on training sets with 300, 200, and 100 images, respectively. The performances of the NHBS-Net model with decreasing training size are the most stable. When the training set sizes are decreased by 25%, 50%, and 75%, the NHBS-Net model's performances drop by 0.59%, 1.21%, and 1.52% in terms of average DSCs, respectively. Fig. 5 shows that the drops in the other CNNs' performance under the training set reduction are noticeable. Moreover, we analyze the sizes of the weight matrix in all methods. The matrix size are 117.9M in Seg-Net, 240.0M in U-Net, 102.7M in FCN, 744.2M in AU-Net, 198.3M in DA-Net, 354.0M in ScSE-Net, and 123.2M in BiSe-Net, respectively. Thus, our NHBS-Net, which has a matrix size of 173.4 M, has an intermediate weight matrix size and gains the best segmentation result. AU-Net, which has approximately 4 times as many parameters as NHBS-Net, has the best performance of the other CNNs.

E. Ablation Experiments

In Table III, we implement segmentation models that contain different parts of the modules mentioned in our NHBS-Net in the 400-image and 100-image trainings, and the results show the average performance for the seven key structures based on three metrics (DSC, HD, and AHD). Table III shows that our NHBS-Net, which contains all modules, has the best performance for both 100-image and 400-image training. The CCOH brings an average DSC improvement of 0.767% for the 400-image model and 1.110% for the 100-image model. The

TABLE III

SEGMENTATION PERFORMANCE SHOWN IN THE ABLATION STUDY OF THE MODELS UNDER 400-IMAGE AND 100-IMAGE TRAINING. THE BASELINE IN OUR NHBS-NET IS THE DILATED RESNET [36] BACKBONE WITH AN OUTPUT HEAD. WE DESIGNED THREE MODULES TO IMPROVE THE SEGMENTATION PERFORMANCE: EDAM WITH ENHANCED CHANNEL ATTENTION(ECA) AND LOCATION ATTENTION MODULE (LAM) WHICH IS INVOLVED IN THE DUAL ATTENTION (DA) MODULE [24], 2-CLASS FFM WITH FIRST-CLASS FFM AND SECOND-CLASS FFM, AND THE CCOH.

Models	400-image training set				100-image training set			
	DSC drop (%)	DSC (%)	HD (pixel)	AHD (pixel)	DSC drop (%)	DSC (%)	HD (pixel)	AHD (pixel)
baseline	↓ 3.155	84.70±5.95	9.92 ±7.57	0.430 ±0.456	↓ 3.479	82.85±6.37	10.48 ±8.30	0.501 ±0.512
baseline+EDAM+ 2-class FFM+CCOH (NHBS-Net)	-	87.85±4.66	8.42±6.30	0.316 ±0.400	-	86.33±5.40	9.63 ±7.18	0.381±0.464
baseline+CCOH	↓ 2.140	85.20±5.25	8.96 ±7.14	0.386 ±0.467	↓ 2.503	83.83±6.32	11.15 ±8.86	0.504 ±0.655
baseline+EDAM	↓ 2.651	85.71±5.39	9.20 ±6.82	0.363 ±0.384	↓ 2.364	83.97±6.10	10.15 ±8.27	0.477 ±0.691
Coordinate Convolution Output Head (CCOH)								
baseline+EDAM+ 2-class FFM	↓ 0.767	87.08±5.03	9.09 ±6.65	0.361 ±0.448	↓ 1.110	85.22±5.71	10.72 ±8.86	0.405 ±0.443
2-class FFM								
baseline+EDAM+CCOH	↓ 1.950	85.90±5.24	9.36 ±6.45	0.376 ±0.406	↓ 1.497	84.83±5.72	12.43 ±12.31	0.477 ±0.608
baseline+EDAM+ first-class FFM+CCOH	↓ 0.903	86.64±5.08	9.05 ±7.50	0.363±0.514	↓ 0.886	85.44±5.48	10.92±9.56	0.413±0.435
baseline+EDAM+ second-class FFM+CCOH	↓ 1.207	86.95±4.85	9.06±7.03	0.357 ±0.458	↓ 1.168	85.16±5.74	10.84±8.50	0.446±0.541
Enhanced Dual Attention Module (EDAM)								
baseline+DA [24]+ 2-class FFM+CCOH	↓ 0.698	87.15±4.77	8.87±7.29	0.338±0.433	↓ 0.917	85.41±5.39	11.39±9.27	0.442±0.487
baseline+LAM [24]+ second-class FFM+CCOH	↓ 1.172	86.68±5.11	9.44±7.58	0.367±0.455	↓ 1.194	85.14±5.71	10.21±8.56	0.433±0.557
baseline+ECA+ second-class FFM+CCOH	↓ 0.563	87.29±5.01	9.58±8.10	0.362 ±0.454	↓ 0.791	85.54±5.54	11.38 ±11.95	0.461 ±0.761
baseline+ second-class FFM+CCOH	↓ 1.531	86.32±5.23	9.25±6.48	0.369±0.409	↓ 1.435	84.89±6.31	10.27 ±9.39	0.429 ±0.525

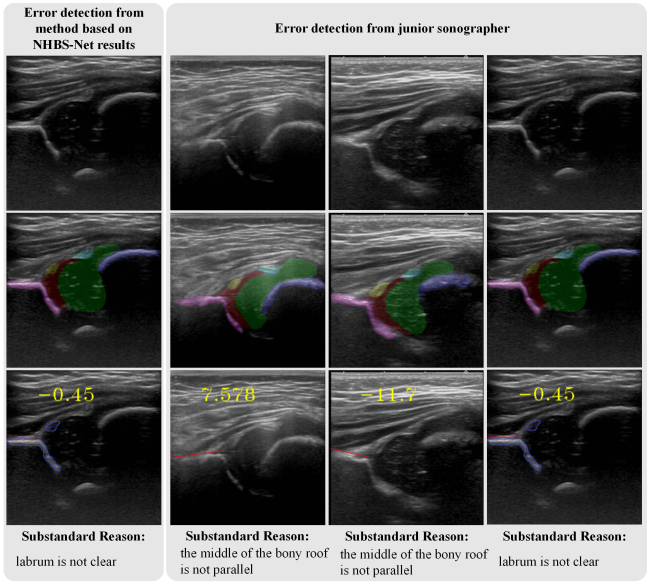

Fig. 6. Some incorrect detection image samples were detected by our method based on NHBS-Net and two healthcare workers. **A.** Incorrect detection using our method in the SPJ dataset. **B.** Incorrect detections by the sonographer in the SPJ dataset.

2-class FFM brings an average DSC increase of 1.950% for the 400-image model and 1.497% for the 100-image model. Furthermore, we explore how the first-class and second-class FFM influence the model performance. In the 400-image model, the first-class FFM brings a 1.207% increase in DSC, and the second-class FFM obtains a 0.903% increase. The 100-image model gains 1.688% and 0.886% increases with the first-class FFM and second-class FFM, respectively.

Additionally, EDAM brings improvements of 1.531% in the 400-image dataset and 1.435% in the 100-image dataset. When the enhanced channel attention part of the dual-channel attention is removed, the performance drops by 1.172% in the 400-image dataset and 1.194% in the 100-image dataset, and removing the location channel results in performance drops of 0.563% in the 400-image dataset and 0.791% in the 100-image dataset. Additionally, we compare the enhanced channel attention module and the channel attention module proposed in [24]. The enhanced channel attention module gains an improvement in DSC, which is 0.698% in the 400-image dataset and 0.917% in the 100-image dataset. For 100-image training, the best DSC performance among the other CNNs is 83.87% obtained by BiSe-Net. In the ablation study, the baseline model using Dilated ResNet obtained an 82.85% in DSC, which is lower than BiSe-Net for 100 training images. The baseline model, with the addition of any of the modules proposed in NHBS-Net, obtained segmentation results equal to or better than those of BiSe-Net. The model performs much better than the other CNNs after adding a combination of modules.

F. Standard Plane Detection Based on NHBS-Net

The key structure segmentation and detection method based on segmentation could help standard plane detection of DDH. In this section, a 50-image quality dataset (the SPJ dataset) is used to test the ability of the detection method based on NHBS-Net. A preliminary user study was conducted on

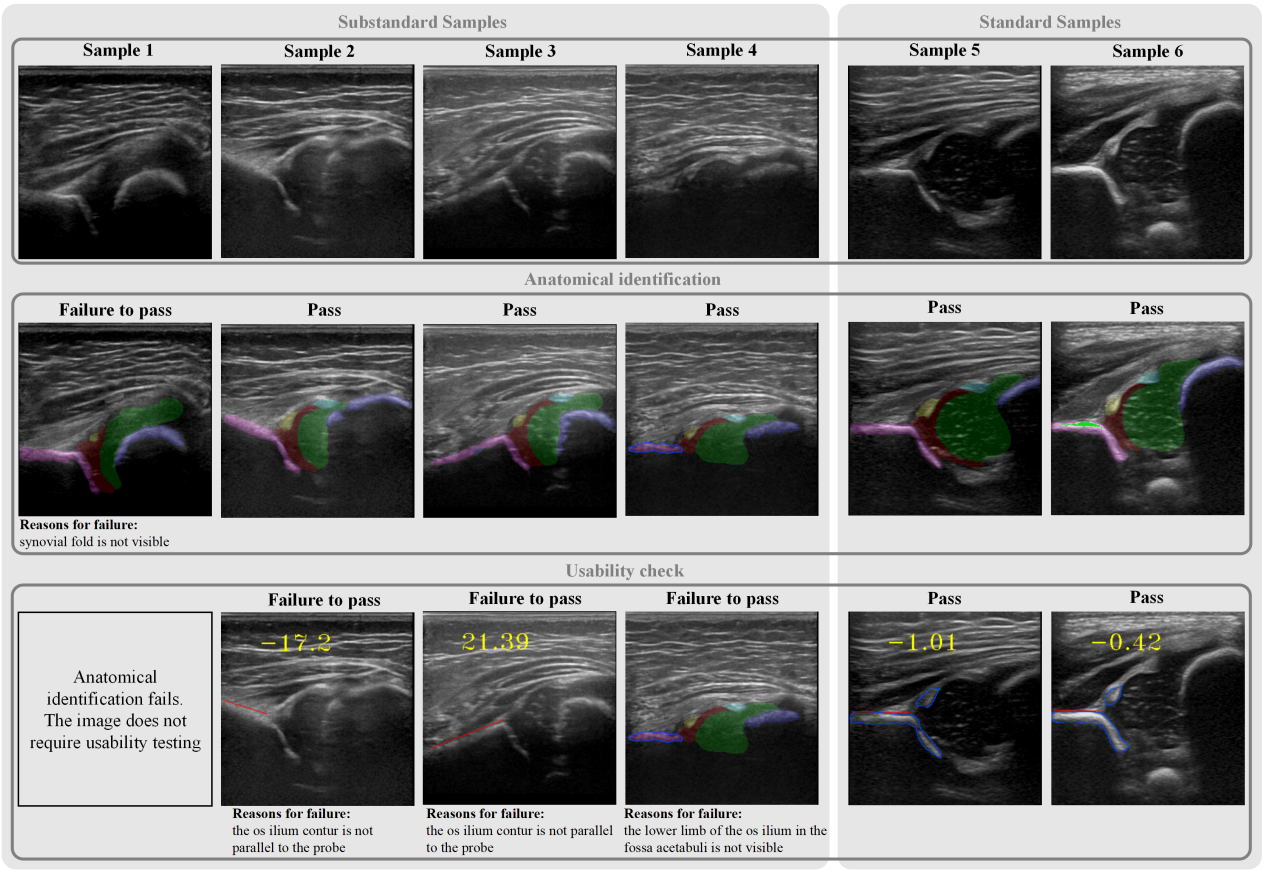

Fig. 7. Some standard and substandard image samples detected by our method based on NHBS-Net.

the SPJ dataset with two healthcare workers using/not using algorithmic aid. The error rate for healthcare worker 1 was 6% (3 images in the 50-image SPJ dataset), while the error rate for healthcare worker 2 was 10% (5 images in the SPJ dataset). Using our automated detection aid, both healthcare worker 1 and 2 achieved error rates of 2% (only one image in the 50-image SPJ dataset). Due to the fact that the reduction in error rate is very preliminary results, only observed in a pilot study of healthcare workers (N=2), we will take more efforts on validating the effectiveness of the proposed techniques in the clinical practice in our future work. In Fig. 6, the error detections of healthcare workers and our method are shown. From the results of our method, the incorrect judgments come from structure clarity errors (Fig. 6 A). The errors of the healthcare worker (junior sonographer) include structure clarity errors and angle deviation in parallelism detection (Fig. 6 B). Some standard and substandard images and visual results of standard plane detection are shown in Fig. 7.

G. Comparisons on the FHU Dataset with the State-of-the-Art CNN Models

Due to the close relationship between ultrasound standardization and clear structure segmentation, incorporating our NHBS-Net into other obstetric ultrasound examinations is helpful. To verify the segmentation ability of our NHBS-Net, we applied the NHBS-Net structure to the FHU dataset, which is a fetal heart ultrasound dataset collected at the Shanghai Jiao Tong University Affiliated Sixth People's Hospital. We trained the models (Seg-Net [31], U-Net [29], FCN [11], AU-Net [17], DA-Net [24], BiSe-Net [25], ScSE-Net [37], and our NHBS-Net) on the FHU training set (76 images, 26 cases) and validated the models' performance on the validation set (28 images, 11 cases). Furthermore, we performed the same operations as on the NHU dataset. The final models, which were the best-performing models on the validation set, were tested on the testing data, including 28 images from 13 cases. The model performance, including the DSC, HD, and AHD of the methods mentioned above, is shown in Table IV. For 5 structures (left atrium, right atrium, left ventricle, right ventricle, and aorta), our NHBS-Net model gained 71.63% in the left ventricle, 71.15% in the right ventricle, 77.66% in the right atrium, 78.96% in the left atrium, and 74.38% in the aorta in terms of DSC. The average DSC for all the structures is 74.75%, which is the best segmentation result among all models.

V. Conclusion and Future Work

Automatic hip joint structure segmentation based on ultrasound images is essential in clinical practice because it can provide visual hints to help sonographers diagnose DDH. Furthermore, it is more helpful for sonographers to obtain segmentation results for seven key structures, which can help in

TABLE IV

SEGMENTATION PERFORMANCE COMPARISON OF OUR NHBS-NET AND OTHER NOVEL CNNS, INCLUDING SEG-NET [31], U-NET [29], FCN [11], AU-NET [17], DA-NET [24], SCSE-NET [37], BISE-NET [25], DEMONSTRATED IN 28-IMAGES FHU TESTING SET UNDER 76-IMAGE TRAINING.

Model name	Average Performance	Five key structures of Heart				
		left ventricle	right ventricle	right atrium	left atrium	aorta
Dice Similiarity Coefficient (DSC) – %						
NHBS-Net	**74.75 ± 25.88**	**71.63 ± 29.10**	**71.15 ± 27.41**	**77.66 ± 27.05**	**78.96 ± 17.87**	**74.38 ± 27.99**
Seg-Net [31]	56.91±27.95	56.77±27.64	56.52±29.23	60.08±22.01	49.59±28.70	61.62±32.16
U-Net [29]	67.61±29.76	69.05±28.21	63.26±30.64	64.92±36.85	73.74±24.10	67.11±29.01
FCN [11]	69.73±29.24	70.46±27.88	69.79±27.87	70.55±33.61	69.87±26.34	67.99±30.52
AU-Net [17]	63.79±29.87	64.15±29.46	65.64±27.03	69.25±27.79	62.31±30.81	57.62±34.27
DA-Net [24]	66.68±29.01	65.26±31.80	70.25±25.61	65.97±31.58	71.46±24.25	60.49±31.82
ScSE-Net [37]	63.44±27.91	67.71±27.55	63.16±30.29	71.83±25.63	70.80±20.98	43.72±35.12
BiSe-Net [25]	69.42±27.14	70.17±28.22	64.17±32.70	71.31±29.57	73.97±17.98	67.46±27.20
Hausdoff Distance (HD) – pixel						
NHBS-Net	**17.13 ± 16.44**	**24.72 ± 25.28**	**18.18 ± 12.07**	17.12 ±17.54	**18.71 ± 18.90**	6.90 ±8.42
Seg-Net [31]	40.57 ±23.49	59.44 ±27.34	44.18 ±27.14	50.17 ±12.66	25.01 ±16.29	24.05 ±34.05
U-Net [29]	29.70 ±30.03	38.71 ±33.03	42.84 ±39.75	17.38 ±16.35	19.39 ±16.19	30.15 ±44.82
FCN [11]	18.13 ±19.06	22.75 ±21.94	26.39 ±31.82	**14.91 ± 15.31**	19.06 ±17.55	7.55 ±8.69
AU-Net [17]	24.91 ±20.05	40.27 ±28.89	26.65 ±21.63	26.76 ±23.32	23.16 ±18.39	7.73 ±8.03
DA-Net [24]	22.80 ±21.45	26.00 ±27.60	24.02 ±17.51	23.07 ±19.69	24.97 ±20.84	15.91 ±21.61
ScSE-Net [37]	33.18 ±22.52	55.63 ±26.18	34.39 ±32.76	26.16 ±17.85	25.63 ±19.48	24.10 ±16.36
BiSe-Net [25]	19.34 ±16.92	28.84 ±26.22	21.85 ±19.95	21.97 ±19.36	18.97 ±16.60	**5.07 ± 2.47**
Average Hausdoff Distance (AHD) – pixel						
NHBS-Net	**2.063 ± 5.154**	3.575 ±10.593	**1.929 ± 3.033**	**1.828 ± 4.982**	2.288 ±5.692	0.695 ±1.468
Seg-Net [31]	6.547 ±8.763	8.873 ±9.597	5.569 ±6.619	6.896 ±6.667	4.411 ±5.897	6.985 ±15.035
U-Net [29]	4.148 ±9.089	4.485 ±7.941	5.682 ±14.151	2.502 ±6.636	**1.387 ± 1.795**	6.685 ±14.921
FCN [11]	2.322 ±5.660	**3.190 ± 7.608**	3.081 ±7.171	2.695 ±8.946	2.170 ±4.084	**0.475 ± 0.492**
AU-Net [17]	3.203 ±5.058	6.944 ±10.063	2.141 ±2.554	3.998 ±8.442	2.336 ±3.648	0.598 ±0.580
DA-Net [24]	3.201 ±6.099	4.585 ±9.493	2.010 ±2.486	4.410 ±9.023	2.152 ±2.739	2.845 ±6.756
ScSE-Net [37]	11.303 ±21.015	8.967 ±18.174	13.544 ±31.317	4.534 ±10.027	2.912 ±4.203	26.557 ±41.353
BiSe-Net [25]	2.375 ±4.734	3.867 ±7.763	2.146 ±3.175	3.366 ±9.066	1.966 ±3.176	0.528 ±0.490

screening standard ultrasound images. Standard plane images need to meet anatomical identification and usability check, which can be helpful in clearly identifying the seven structures and making visibility and parallelism judgments based on structure segmentation. However, due to the limitations of the scale of data collection and labeling complexity, robust segmentation models that can segment the seven hip joint structures are still being explored. Our work illustrates NHBS-Net, the first model for segmenting the seven key neonatal hip joint structures. The experimental results show that NHBS-Net, using a 400-image training set for development, has excellent segmentation performance on our test dataset. Due to the time-consuming and laborious work of collecting a large dataset with annotations, we compared the model performance impact as the training set size decreased. Stable segmentation performance is achieved by our NHBS-Net model with a decreasing size in the training dataset. Due to the generalization of ultrasound image standardization, our NHBS-Net can also be applied in other ultrasound segmentation work to clarify structures, ensuring standard image acquisition. We develop a standard plane detection method based on segmentation of the seven structures and implement the method on a standard plane judgment dataset (the SPJ dataset). We also implement our NHBS-Net in another ultrasound dataset, the FHU dataset, and NHBS-Net obtains excellent results on the FHU dataset; the performance exceeds that of the other CNNs by 5%.

Although a very preliminary user study was conducted to compare the detection error rate of two healthcare workers by using/not using our method, the randomized controlled trial (RCT) is necessary to be conducted for further validating effectiveness and efficiency of the proposed technique. In the future, we will focus on the development of a segmentation-based standard plane detection model and its applications. Furthermore, we will explore segmentation-based classification methods for DDH severity measurement and classification.

REFERENCES

[1] J. C. Jackson, M. M. Runge, and N. S. Nye, "Common questions about developmental dysplasia of the hip," *Am. Fam. Physician*, vol. 90, no. 12, pp. 843–850, 2014.

[2] O. Furnes, S. A. Lie, B. Espehaug, S. E. Vollset, L. B. Engesaeter, and L. I. Havelin, "Hip disease and the prognosis of total hip replacements," *J. Bone. Joint. Surg.*, vol. 83-B, no. 4, pp. 579–586, 2001.

[3] R. Graf, "The diagnosis of congenital hip-joint dislocation by the ultrasonic combound treatment," *Arch. Orthop. Traumat. Surg.*, vol. 97, pp. 117–133, 1980.

[4] R. Graf, M. Mohajer, and F. Plattner, "Hip sonography update. Quality-management, catastrophes - tips and tricks," *Med. Ultrason.*, vol. 15, no. 4, pp. 299–303, 2013.

[5] N. Quader, A. Hodgson, K. Mulpuri, T. Savage, and R. Abugharbieh, "Automatic assessment of developmental dysplasia of the hip," in *Proc. ISBI*, 2015, pp. 13–16.

[6] P. Pandey, N. Quader, K. Mulpuri, P. Guy, R. Garbi, and A. J. Hodgson, "Shadow peak: Accurate real-time bone segmentation for ultrasound and developmental dysplasia of the hip," in *The 19th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery (CAOS)*, vol. 3, 2019, pp. 301–305.

[7] H. B. Sezer and A. Sezer, "Automatic segmentation and classification of neonatal hips according to Grafs sonographic method: A computer-aided diagnosis system," *Appl. Soft. Comput.*, vol. 82, p. 105516, 2019.

[8] G. Wu, H. Jia, Q. Wang, and D. Shen, "Sharpmean: Groupwise registration guided by sharp mean image and tree-based registration," *NeuroImage*, vol. 56, no. 4, pp. 1968–1981, 2011.

[9] H. Jia, P. Yap, and D. Shen, "Iterative multi-atlas-based multi-image segmentation with tree-based registration," *NeuroImage*, vol. 59, no. 1, pp. 422–430, 2012.

[10] Y. Zhan, M. D. Feldman, J. E. Tomaszeweski, C. Davatzikos, and D. Shen, "Registering histological and MR images of prostate for image-based cancer detection," in *Proc. MICCAI*, vol. 4191, 2006, pp. 620–628.

[11] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 833–851.

[13] T.-C. Chiang, Y.-S. Huang, R.-T. Chen, C.-S. Huang, and R.-F. Chang, "Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation," *IEEE Trans. Med. Imaging*, vol. 38, no. 1, pp. 240–249, 2019.

[14] X. Yang, L. Yu, S. Li, H. Wen, D. Luo, C. Bian, J. Qin, D. Ni, and P.-A. Heng, "Towards automated semantic segmentation in prenatal volumetric ultrasound," *IEEE Trans. Med. Imaging*, vol. 38, no. 1, pp. 180–193, 2019.

[15] Y. Wang, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P.-A. Heng, T. Wang, and D. Ni, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imaging*, vol. 38, no. 12, pp. 2768–2778, 2019.

[16] R. Zhou, F. Guo, M. R. Azarpazhooh, J. D. Spence, E. Ukwatta, M. Ding, and A. Fenster, "A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3D ultrasound images," *IEEE Trans. Med. Imaging*, vol. 39, no. 9, pp. 2844–2855, 2020.

[17] H. Sun, C. Li, B. Liu, Z. Liu, M. Wang, H. Zheng, D. D. Feng, and S. Wang, "AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms," *Phys. Med. Biol.*, vol. 65, p. 055005, 2020.

[18] C. Kaul, S. Manandhar, and N. Pears, "Focusnet: An attention-based fully convolutional network for medical image segmentation," in *Proc. ISBI*, 2019, pp. 455–458.

[19] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Inform.*, pp. 1–1, 2020.

[20] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2289–2301, 2020.

[21] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" in *International Conference on Learning Representations*, 2020.

[22] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution." in *Proc. NeurIPS*, 2018, pp. 9605–9616.

[23] T. Song, J. Chen, X. Luo, Y. Huang, X. Liu, N. Huang, Y. Chen, Z. Ye, H. Sheng, S. Zhang, and G. Wang, "CPM-Net: A 3D center-points matching network for pulmonary nodule detection in CT scans," in *Proc. MICCAI*, vol. 12266, 2020, pp. 550–559.

[24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. CVPR*, 2019, pp. 3141–3149.

[25] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Song, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, 2018, pp. 334–349.

[26] A. R. Hareendranathan, D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko, "Toward automatic diagnosis of hip dysplasia from 2D ultrasound," in *Proc. ISBI*, 2017, pp. 982–985.

[27] Z. Zhang, M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand, and J. L. Jaremko, "End-to-end detection-segmentation network with ROI convolution," in *Proc. ISBI*, 2018, pp. 1509–1512.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[30] H. El-Hariri, K. Mulpuri, A. Hodgson, and R. Garbi, "Comparative evaluation of hand-engineered and deep-learned features for neonatal hip bone segmentation in ultrasound," in *Proc. MICCAI*, 2019, pp. 12–20.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 936–944.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[36] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," in *Proc. CVPR*, 2017, pp. 636–644.

[37] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. MICCAI*, vol. 11070, 2018, pp. 421–429.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[39] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, 2018, pp. 7794–7803.

[40] D. Novotný, S. Albanie, D. Larlus, and A. Vedaldi, "Semi-convolutional operators for instance segmentation," in *Proc. ECCV*, vol. 11205, 2018, pp. 89–105.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.

[42] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[43] P. F. Raudaschl *et al.*, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015," *Med. Phys.*, vol. 44, no. 5, pp. 2020–2036, 2017.

[44] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993.

[45] D. Ni, X. Yang, X. Chen, C. T. Chin, S. Chen, P. A. Heng, S. Li, J. Qin, and T. Wang, "Standard plane localization in ultrasound by radial component model and selective search," *Ultrasound in Medicine and Biology*, vol. 40, no. 11, pp. 2728–2742, 2014.

[46] D. Ni, T. Li, X. Yang, J. Qin, S. Li, C.-T. Chin, S. Ouyang, T. Wang, and S. Chen, "Selective search and sequential detection for standard plane localization in ultrasound," in *Proceedings of the 5th International Workshop on Abdominal Imaging. Computation and Clinical Applications - Volume 8198*, 2013, pp. 203–211.