Ultrasound Standard Plane Detection Using a Composite Neural Network Framework

Hao Chen, *Student Member, IEEE*, Lingyun Wu, Qi Dou, *Student Member, IEEE*, Jing Qin, *Member, IEEE*, Shengli Li, Jie-Zhi Cheng, *Member, IEEE*, Dong Ni, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*—Ultrasound (US) imaging is a widely used screening tool for obstetric examination and diagnosis. Accurate acquisition of fetal standard planes with key anatomical structures is very crucial for substantial biometric measurement and diagnosis. However, the standard plane acquisition is a labor-intensive task and requires operator equipped with a thorough knowledge of fetal anatomy. Therefore, automatic approaches are highly demanded in clinical practice to alleviate the workload and boost the examination efficiency. The automatic detection of standard planes from US videos remains a challenging problem due to the high intraclass and low interclass variations of standard planes, and the relatively low image quality. Unlike previous studies which were specifically designed for individual anatomical standard planes, respectively, we present a general framework for the automatic identification of different standard planes from US videos. Distinct from conventional way that devises hand-crafted visual features for detection, our framework explores in- and between-plane feature learning with a novel composite framework of the convolutional and recurrent neural networks. To further address the issue of limited training data, a multitask learning framework is implemented to exploit common knowledge across detection tasks of distinctive standard planes for the augmentation of feature learning. Extensive experiments have been conducted on hundreds of US fetus videos to corroborate the better efficacy of the proposed framework on the difficult standard plane detection problem.**

Manuscript received December 6, 2016; revised March 12, 2017; accepted March 16, 2017. This work was supported in part by the National Basic Research Program of China, 973 Program under Project 2015CB351706, in part by the National Natural Science Foundation of China under Grant 61571304, Grant 61501305, and Grant 61233012, and in part by the Research Grants Council of Hong Kong Special Administrative Region under Grant CUHK 14202514. This paper was recommended by Associate Editor M. Shin. *(Corresponding authors: Jie-Zhi Cheng; Dong Ni.)*

H. Chen and Q. Dou are with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong (e-mail: jackie.haochen@gmail.com).

L. Wu, J.-Z. Cheng, and D. Ni are with the School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen University, Shenzhen 518060, China (e-mail: jzcheng@szu.edu.cn; nidong@szu.edu.cn).

J. Qin is with the School of Nursing, Hong Kong Polytechnic University, Hong Kong.

S. Li is with the Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital, Nanfang Medical University, Shenzhen 518000, China.

P.-A. Heng is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, and also with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518052, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2685080

***Index Terms*—Convolutional neural network (CNN), deep learning, knowledge transfer, recurrent neural network (RNN), standard plane, ultrasound (US).**

I. Introduction

ULTRASOUND (US) is a widely used obstetric examination tool for its advantages of low cost, mobility, and the capability of real time imaging [1], [2]. In general, the clinical obstetric US examination involves the procedures of manual scanning, standard plane selection, biometric measurement, and diagnosis [3]. Particularly, the accurate acquisition and selection of the US planes that can clearly depict the key anatomic structures of fetus is very crucial for the subsequent biometric measurement and diagnosis. For example, the prebirth weight of baby can be estimated from the US measurements of head circumference, biparietal diameter, abdominal circumference, and femur length. Therefore, the selection of US planes that can depict the corresponding organs with good quality will be very important for the accurate estimation of fetus weight [4], [5]. In terms of diagnostic purpose, the US views that can visualize the detailed facial and cardiac structures of fetus deem to be very important for the timely prenatal diagnosis of facial dysmorphism and congenital heart diseases. These US planes that can depict key anatomic structures clearly for either biometric measurement or disease diagnosis are generally recommended by professional organizations for the standard fetal US examination and are often denoted as US standard planes [6]–[9].

In clinical practice, the US standard plane is commonly acquired by hand with laborious maneuver of the probe for searching the desirable view that can concurrently present the key anatomical structures, see Fig. 1. Specifically, three standard planes: 1) fetal abdominal standard plane (FASP); 2) fetal face axial standard plane (FFASP); and 3) fetal four-chamber view standard plane (FFVSP) of heart are shown in Fig. 1. The FFASP is determined with the presence of three key organs of: 1) nose bone; 2) lens; and 3) eyes in the US view, whereas the FASP is expected to include stomach bubble (SB), umbilical vein (UV), and spine (SP). The definition of FFVSP is the US plane that can clearly visualize five key cardiac structures of: 1) left atrium; 2) right atrium; 3) left ventricle; 4) right ventricle; and 5) descending aorta in the same image. The FASP can be used for the estimation of fetal weight, while the FFASP and FFVSP can be informative for the diagnosis of facial dysmorphism and congenital heart diseases, respectively.

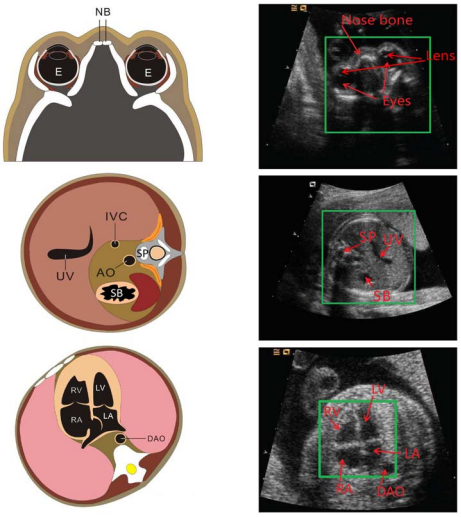

Fig. 1. Illustration of different fetal standard planes for FFASP, FASP, and FFVSP, respectively (left column illustrates the anatomical structures, right column illustrates the corresponding US image examples, and the green rectangles denote the ROI).

Since the clinically needed US standard planes can be very diverse and the overall number of planes can be several dozens for a thorough examination [10], it usually takes around tens of minutes or more to acquire and define the US standard planes, even for a very experienced obstetrician. Therefore, the selection of necessary US standard planes can be one of the most time consuming procedure in the obstetric examination. On the other aspect, the process of acquisition and selection of the correct US standard planes requires the operator being proficient in maternal and fetal anatomy and highly depends on operator's experience. As a consequence, it would be very challenging for an inexperienced operator or novice to fulfill the whole task of US standard plane acquisition. Meanwhile, since the standard plane acquisition is a knowledge-intensive task and required planes are very diverse, the learning curve of this procedure can be very long [11]. In such a case, the manpower shortage can be expected in highly populated regions as the training of a ready operator for the US fetal examination can be costly and take a long time. Motivated by the aforementioned issues, the computerized scheme with automatic plane detection and selection capability will be highly welcome to alleviate the routinely obstetric workload [12] and address the issues of medical manpower shortage on underserved populations and areas [11]. The computer-aided scheme can also help to facilitate the training of medical novices with computerized feedback from a score-based quality control system [13].

The topic of computer-aided US frame detection and selection is relatively new and has recently received more and more attention in these years [8], [12], [14]–[18]. The computerized scheme can help to lower down the operator dependency in US scanning and improve the efficiency of post-processing procedures with automatic mechanisms. Kwitt *et al.* [11] developed a template-based method equipped with dynamic texture model to retrieve frames containing key structures from US video. The efficacy of the template-based method was merely verified in phantom studies, and hence, the applicability to the real data may need to be further explored. In the obstetric application, quite a few computerized methods had also been proposed to identify specific standard planes from freehand US videos. Zhang *et al.* [8] adopted the cascade AdaBoost to locate the plane with gestational sac. To automatically select the FASP from the US video, Ni *et al.* [12] used the radial component descriptor to encode the spatial co-presence relation of the SB, UV, and SP to retrieve the target plane. Generally speaking, most previous methods have to find out useful features and exploit the mathematical and spatial priors for the detection of each specific US plane. In such a case, the detection method designed for one standard plane, e.g., FASP, may not be easily generalized to another standard plane, say FFASP.

By and large, the challenges of developing the detection algorithm for US standard planes can be summarized in fourfold. First, the US standard plane often has high intraclass appearance variation caused by various factors like imaging artifacts of acoustic shadows and speckles, deformation of soft tissues, fetal development, transducer poses [15], [19]–[21], etc. Second, the key anatomical structures in the standard plane may possibly appear similar to other structures. For instance, shadows, the abdominal aorta and the inferior vena cava are often mistakenly identified as the SB or UV in the FASP of Fig. 1, as the shape and echogenicity of these structures resemble to each other. Accordingly, even for experienced obstetricians, the plane selection results can be possibly misled by the low interclass variation. The third challenge lies in that the available US fetus training image data and expert annotations are significantly more limited and less accessible than the image data for many computer vision problems. To obtain the US fetus data, it has to get the local institutional review board (IRB) approval and consent from subjects. Meanwhile, the annotation on the US standard planes from long US fetus videos requires professional obstetric knowledge and is a very time consuming task. With limited training data and annotation, the capability of any US standard plane method based on machine learning will be constrained. The potential over-fitting issue may also be difficult to avoid. The fourth challenge consists in that the US fetus standard planes can be very diverse for their own diagnostic purposes, see Fig. 1. In such a case, it will be very hard to devise a general method that can retrieve multiple standard planes from US fetus videos. These four challenges will impose great difficulty on any off-the-shelf pattern recognition techniques, e.g., the template-based [11], geometrical shape-based [12], and feature-based methods [15], [19], and hence, the algorithm for each standard plane may need to be specifically designed.

The deep learning techniques have made breakthroughs in the field of computer vision [22]–[25] and medical image computing [26]–[32]. Instead of elaboration on hand-crafted features on each respective problem in the conventional pattern recognition pipeline, the deep learning techniques are able to automatically discover important features and exploit the feature relation from training data [33], [34]. However, the deep learning techniques may demand a large number of training data, which is usually not feasible in medical image analysis problems, to construct an effective model. To address the issue

of training data size, the transfer learning scheme has recently been introduced into the deep learning techniques, particularly with the deep convolutional neural networks (CNNs), to leverage the knowledge across different domains [35]–[37]. Specifically, in the application of US fetus standard plane detection, Chen *et al.* [38] exploited to transfer the knowledge from the natural scene images toward the domain of fetus for the identification of FASP with the CNN model. The experimental results suggested that the low level image cues like corner, edges, etc., learned from the natural scene domain can serve as good network initialization for CNN to effectively boost FASP detection performance than the random initialization setting. Although relatively satisfactory performance had been achieved with knowledge-transferred CNN scheme in [38], the gap between the natural scene domain and the fetus US domain remains significant. Accordingly, the performance improvement may be thus limited. Meanwhile, the study of [38] only considered the image cues within single plane, which may not be sufficient to address the high intraclass and low interclass variation issues. As an extension of our previous work [39], in this paper, we will explore the interframe contextual clues, which are very informative for human experts during the manual screening, for the US standard plane detection problem.

To address the four challenges discussed above, this paper attempts to leverage the framework of multitask learning, deep learning technique, and the sequence learning model (RNN) to detect three standard planes, i.e., FASP, FFASP, and FFVSP, from US fetus videos. Specifically, we treat the detection of the three standard planes as three individual tasks and jointly learn the spatial features with the deep CNN. The shared spatial features across the three tasks extracted from individual frame are further transferred to the RNN for the modeling of temporal relation. The multitask learning framework aims to uncover the common knowledge shared across different tasks. With such a framework, the training data on each individual task can be helpful to other tasks, and hence, the demand on large data size for all tasks can be potentially eased. The training of the deep CNN is based on the multitask learning framework to identify the useful common in-plane spatial features at the supervised learning phase. With the consideration of learning the three detection tasks in the same architecture, the generalization capability of the constructed deep CNN can be thus augmented and the issues of low interclass and high interclass variations can also be handled properly. The RNN [40]–[42] had been widely applied to address many machine learning problems for various sequential data, e.g., speech recognition [43], video recognition [44], [45], and machine translation [46], with promising results. In this paper, we specifically exploit the long short-term memory (LSTM) model to harness the interframe contexts. The LSTM model has a good capability to solve issues of exploding or vanishing gradients that could be possibly caused by temporal signal drops, serious noise corruption, and occlusion [47], [48]. The training of the LSTM model is based on the extracted features from the multitask deep CNN. Since the contextual cues are also commonly used by medical experts in the clinical US scanning and plane selection, the modeling of interframe contexts may be helpful to tackle the issue of low interclass variation for better detection performance.

The performance of the proposed multitask deep and temporal learning framework will be evaluated with extensive experiments by comparing our performance with other state-of-the-art methods in the literature. The outperformance of the proposed method over other baseline methods corroborates the efficacy of multitask learning and the exploit of temporal features on this new US standard plane detection problem. Since the proposed method does not explicitly elaborate on the feature design, it is also easy to apply our multitask deep and temporal learning framework for the detection of other standard US planes.

The remainder of this paper is organized as follows. Section II describes the proposed method in details. Experimental results are evaluated qualitatively and quantitatively in Section III. Section IV discusses the advantages and disadvantages of our proposed method, as well as future research directions. Finally, the conclusions are drawn in Section V.

II. Method

The left part of Fig. 2 illustrates the overview of the proposed model, which is a composite neural network framework with the specialized deep CNN and RNN to exploit the in- and between-plane features from fetus US videos. The composite neural network is denoted as T-RNN throughout this paper for short. The deep CNN model of the T-RNN framework aims to uncover useful spatial features from individual US planes. To address the issue of limited training data, the multitask learning is implemented for the training of the deep CNN models by treating the detection of FASP, FFVSP, and FFASP as three individual tasks. The goal of multitask learning is to leverage the limited training data of each detection task for better model generalization and avoidance of potential over-fitting problem. Comparing to the large domain gap between the natural scene images and US fetus images [38], the training data of the three detection tasks are of the same image modality and relatively relevant. Therefore, the common knowledge shared by the three detection tasks may be more easily explored by the CNN model, and would be served as a more reliable basis for the task-oriented fine-tuning. Based on the in-plane knowledge of the CNN models learned with the multitask learning, the between-plane relation is further exploited with the specific RNN of LSTM model. The complex contextual knowledge discovered by the LSTM model will help to deal with the issues of low interclass variation for the boosting of detection capability.

The whole T-RNN framework is realized in three major steps. First, a regions of interest (ROI) classifier is *jointly* trained with CNN models, named as J-CNN, across three detection tasks of FASP, FFVSP, and FFASP. The ROI classifier of J-CNN models is expected to locate the informative regions in each US plane of the three detection tasks. The features extracted from the identified ROI at each frame by the J-CNN model are further forwarded to the LSTM model that is imparted with the between-plane knowledge to yield

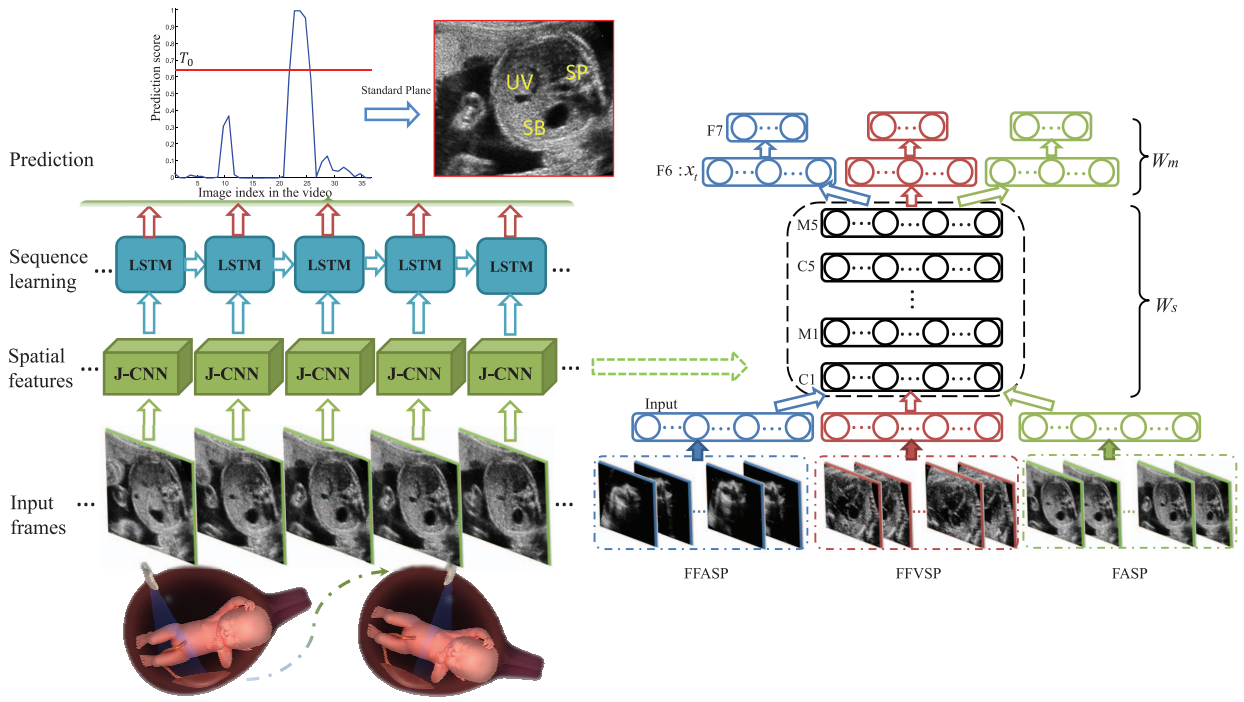

Fig. 2. Left: overview of the proposed T-RNN model. Right: architecture of the proposed J-CNN model.

TABLE I
ARCHITECTURE OF J-CNN MODEL

Layer	Feature maps	Kernel size	Stride
input	227x227x1	-	-
C1	55x55x24	11	4
M1	27x27x24	3	2
C2	14x14x24	5	2
M2	7x7x24	3	2
C3	7x7x24	3	1
C4	7x7x24	3	1
C5	7x7x24	3	1
M5	3x3x24	3	2
F6	100	-	-
F7	2	-	-

the corresponding task prediction scores on each US frame. Finally, the score of each frame is further inferred by averaging all prediction scores from the LSTM model. A US plane will be identified as the standard plane when the inferred score is larger than a defined threshold T_0.

A. Joint Learning Across Multitasks

The basic structure of CNN is composed of several pairs of alternating convolutional (C) and max-pooling (M) layers, followed by fully connected (F) layers [49]. Previous studies have suggested that the knowledge learned from one task domain via CNN can benefit the training of another task domain where annotated data are limited [35], [36], [38], [50], [51]. Therefore, the CNN model can be very suitable for multitask learning. Specifically, a joint learning scheme with CNN across multiple detection tasks of US standard planes is carried out, as illustrated in the right part of Fig. 2. The matrix W_s represents the synaptic parameters of layers from $C1$ to $M5$ and can be adjusted in training process of the CNN model. Via the co-training process from the annotated data of the FASP, FFVSP, and FFASP, the common knowledge across the three distinctive tasks can be further encoded in the matrix W_s. The W_m ($m = 1, 2$, and 3 represents the task of FFASP, FFVSP, and FASP, respectively) stands for the synaptic parameters of $F6$ and $F7$ layers to learn the task-specific knowledge at the supervised training of the CNN models. The whole learning problem is then formulated as a cost minimization process of the joint max-margin loss function $\mathcal{L}_1$

$$\mathcal{L}_1 = \frac{\lambda}{2}\left(\sum_m \|W_m\|_2^2 + \|W_s\|_2^2\right) + \sum_m \sum_k \max\left(0, 1 - y_{mk} F_m\left(f_{mk}^s; W_m\right)\right)^2 \quad (1)$$

$$f_{mk}^s = F_s(I_{mk}; W_s) \quad (2)$$

where the first component of $\mathcal{L}_1$ is the regularization penalty term and the second component is the data loss term. The cost minimization can be realized by adjusting the synaptic matrices of W_s and W_m. The importance weighting between the two terms in (1) is controlled by the hyper-parameter λ, which is empirically defined as 1.0 throughout this paper. In (2), the function F_s indicates the common feature function specified by W_s across the three tasks, whereas the function F_m is the task-specific discriminant function controlled by matrix W_m. The I_{mk} in (2) stands for the kth image plane with respect to the mth task, and the f_{mk}^s is the output of the function F_s, i.e., the neuron activations of $M5$ layer. The $y_{mk} \in \{-1, 1\}$ specifies the corresponding ground truth label for the input frame I_{mk}. The detailed architecture configuration of the J-CNN models in this paper can be found in Table I, where padding and nonlinear activation layers are not shown for simplified presentation. Meanwhile, the rectified linear units are implemented

in the nonlinear activation layer [52] and the dropout strategy is employed in the fully connected layers for better generalization capability [53]. The learning rate is set as 0.01 initially and gradually decreased by factor of 10, whenever the training loss stops to decrease. The constructed J-CNN models can help to manifest the informative ROIs with respect to each task and the corresponding extracted features will be fed into the latter LSTM model for further processing.

B. US Standard Plane Detection via T-RNN

During the clinical US fetal examination, the contextual cues between two consecutive scanning frames are intuitively used by the operator for the searching of anatomical targets, as the in-plane visual cues sometimes may not be sufficient to support the clinical judgement. Motivated by this, the special RNN model, i.e., the LSTM [47], is adopted here to exploit the between-plane cues from the recorded US fetus videos. The training of the LSTM model is based on the manifested in-plane ROIs from the J-CNN model. Because the J-CNN model can filter out most irrelevant image cues to the three detection tasks, the LSTM model can further focus on the polished task-related ROI for more efficient and effective establishment of contextual relations between US planes.

Given the input frame I_{mk}, the probability map of the ROI is computed by the J-CNN model with the sliding window technique. Specifically, for robustness of computation, each subimage by the sliding window from the original image is augmented into ten input samples by cropping the patches of its center and four corners, as well as the corresponding mirrored five patches [38]. The final score of each sliding window subimage can then be defined with the averaged J-CNN score over its 10 varied replications. With the robust sliding window scheme, the center of the final ROI identified by J-CNN can be regarded as the location with maximal value in the computed probability score map. Following that, the features from the penultimate layer (i.e., the activations of $F6$ layer) of the J-CNN model are extracted from the estimated ROI of each frame as the inputs of the LSTM model. A preprocessing of the US videos is implemented to facilitate the training of LSTM model. Specifically, the long US videos are clipped into shorter montages of fixed T frames. Accordingly, the input video can be thus treated as consecutive samples of montages. Each montage is denoted by a sequential feature vector: $\mathbf{x} = \{x_1, \ldots, x_t, \ldots, x_T\}$ and $x_t \in \mathbb{R}^q$ ($q = 100$ in our experiments) with the corresponding label vector of $\mathbf{y} = \{y_1, \ldots, y_t, \ldots, y_T\}$, where $y_t \in \{0, 1\}$. It is worth noting that the consecutive clipped montages share overlapping US frames for the robustness of computation.

In the traditional RNN, the back-propagation algorithm is commonly adopted for the training. However, the back-propagation algorithm may fall short of dealing with the vanishing or exploding gradients [47], [48], and thus could be sensitive to noisy or corruption in the data sequence. The LSTM model on the other hand is able to tackle this problem by incorporating the so-called memory cells into the network architecture. The memory cell equips the network with better abilities to find and exploit long range context with the arrival

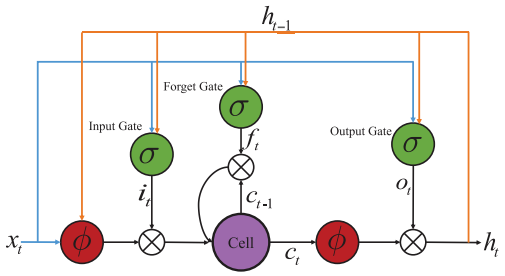

Fig. 3. Illustration of LSTM model.

of sequential inputs [43], hence, it endows the LSTM model the capability and flexibility on handling the intermittent noise, data corruption and error. With these advantages, the LSTM model will be quite suitable for the processing of US fetus videos, where the image quality of some frames can possibly be very bad and not informative.

A basic architecture of LSTM model can be constituted with units of input gate, memory cell wired with self-recurrent connection, forget gate and output gate, see Fig. 3 for illustration. Specifically, the element-wise nonlinear functions shown in Fig. 3 can be either the sigmoid function in the form of $\sigma(x) = [1/(1 + e^{-x})]$ or the hyperbolic tangent function, $\phi(x) = [(e^x - e^{-x})/(e^x + e^{-x})]$, that can squash the range of input x into the respective range of [0, 1] and [−1, 1]. The gates serve to modulate the interactions between the memory cell c_t and its environment [54], [55]. The input gate i_t can control incoming input x_t whether to alter the state of the memory cell or block it instead. The output gate o_t is in charge of the memory cell state to have an effect on hidden neurons or not. The forget gate f_t can modulate the self-recurrent connection of the memory cell to steer the memory cell whether to remember or forget the previous state c_{t-1}. All the gates and memory cell have the same vector size with hidden state $h_t \in \mathbb{R}^H$ (H is the number of hidden units). The update mechanisms of the gates and memory cells can be realized with the following equations:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \tag{3}$$

where $h_0 = 0$, and all W denote the weighting matrices. For examples, W_{xi} is the input-input gate matrix, whereas W_{hi} is the matrix of hidden-input gate. In the (3), all b stand for the bias terms with respect to each unit, and the operator $\odot$ represents the element-wise multiplication. The final predictions can be obtained by feeding h_t into a softmax classification layer over the three tasks. Thus, the parameters θ (including all W and b) of the LSTM model can be trained by minimizing the negative logarithm loss function $\mathcal{L}_2$ with stochastic gradient descent method [56]. The $\mathcal{L}_2$ is defined as

$$\mathcal{L}_2 = -\sum_{n=1}^{N}\sum_{t=1}^{T} \log p_n(y_t|x_t, h_{t-1}; \theta) \tag{4}$$

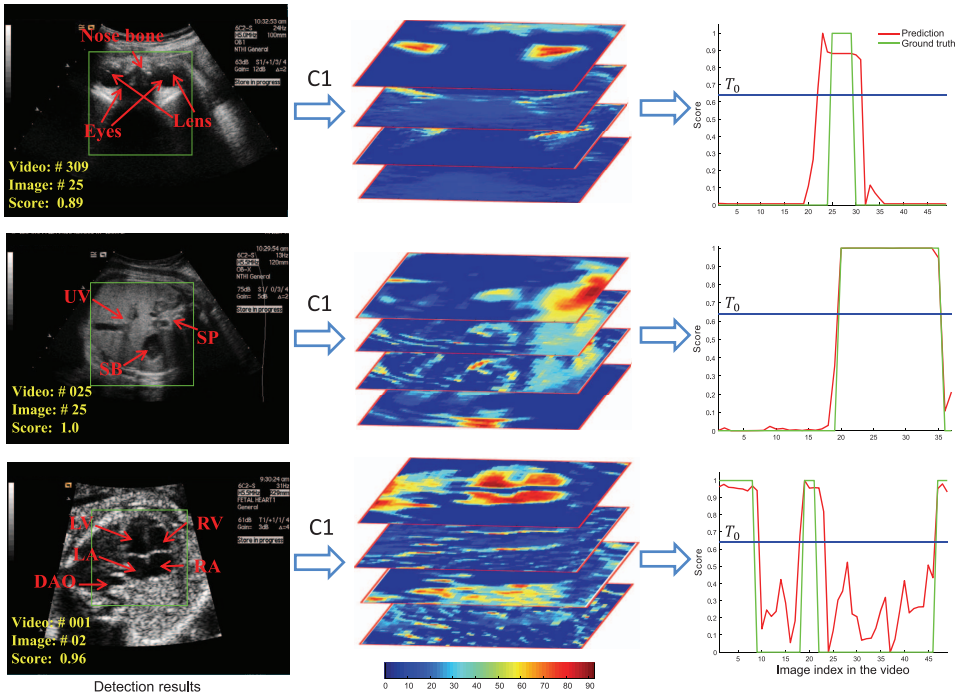

Fig. 4. Left: typical US standard plane detection results. Middle: several feature maps of ROIs in C1 layer. Right: sequenced predictions in the video.

where N is the total number of the clipped montages, and $p_n(y_t|x_t, h_{t-1}; \theta)$ is the correctly predicted probability function for tth frame of one training montage, given the current input x_t and previous hidden state h_{t-1}.

III. Experiments and Results

A. Materials

All the US images and videos involved in this paper were acquired from the Shenzhen Maternal and Child Healthcare Hospital during September 2011 to February 2013. The study protocol was reviewed and approved by the ethics committee of the same institution. Meanwhile, all participating subjects agreed the data usage for scientific research and relevant algorithm development. The US videos were recorded with conventional hand-held 2-D US probe on pregnant women in the supine position, by following the standard obstetric examination protocol. All US videos were acquired with a Siemens Acuson Sequoia 512 US scanner, and the fetal gestational age of all subjects ranges from 18 to 40 weeks. Each video was obtained from one subject with 17–48 US frames for the purpose of searching one US standard plane. More specifically, one US video can be recorded from the region of either fetal face, abdomen, or chest to enclose the respective FFASP, FASP, or FFVSP. The ground truths of videos were manually annotated by an experienced obstetrician with more than five years of clinical experience.

For the training of the ROI classifier with J-CNN, the training samples with respect to FASP, FFASP and FFVSP were drawn from respective 300 US videos. Therefore, there are totally 900 US videos in which each of them exclusively contains one type of the three standard planes. For the performance evaluation, the tasks of FASP and FFASP are tested with 219 videos and 52 videos, respectively, whereas the testing data for the FFVSP task are 60 videos. The overall involved testing US images for the FASP and FFASP are 8718 and 2278, respectively, and the number of US images for the FFVSP is 2252. All the training and testing data were collected by following the rigorous scanning protocol for quality assurance. Details of the used US dataset in this paper can be found in Table II. In summary, there are a total of 1231 US videos for the training and testing of the proposed T-RNN, whereas the overall number of involved images is 50 624. To the best of our knowledge, this is the largest real clinical dataset available for US standard plane detection study.

TABLE II
Details of US Dataset

US dataset	FASP	FFASP	FFVSP
Train (video/frame)	300/11,942	300/13,091	300/12,343
Test (video/frame)	219/8,718	52/2,278	60/2,252

B. Visualization of Intermediate Results

To give insight on the interaction between the models of J-CNN and LSTM during the processing of US fetus videos, Fig. 4 demonstrates the feature maps of J-CNN for the video frames and the task prediction result on each frame by the LSTM. The left column of Fig. 4 shows the final detection results of three US standard planes by the proposed method T-RNN. It can be observed that all identified standard planes by our algorithm can clearly depict the corresponding key anatomic structures. The predicted scores by the LSTM model of the three identified planes in the left column of Fig. 4 are above the threshold T_0 (determined by testing on a set

TABLE III
RESULTS OF STANDARD PLANE DETECTION ON US IMAGES

Method	FASP				FFASP				FFVSP			
	A	P	R	$F1$	A	P	R	$F1$	A	P	R	$F1$
T-RNN	**0.908**	**0.748**	**0.747**	**0.747**	**0.867**	**0.634**	**0.598**	**0.615**	**0.867**	**0.770**	**0.612**	**0.682**
J-CNN	0.902	0.729	0.739	0.734	0.854	0.605	0.513	0.555	0.835	0.718	0.611	0.660
T-CNN [38]	0.896	0.714	0.710	0.712	0.847	0.582	0.503	0.535	0.831	0.708	0.606	0.653
R-CNN [38]	0.857	0.594	0.681	0.635	0.831	0.530	0.443	0.482	0.826	0.688	0.608	0.646
RVD [12]	0.833	0.532	0.693	0.602	-	-	-	-	-	-	-	-

of samples from the training set in our experiments). To further provide the visual assessment of the detection efficacy of key anatomical structures by the J-CNN, the middle column of Fig. 4 lists the C1 feature maps [57] of the US frame by our J-CNN model. Specifically, it can be found that the regions with large responses of C1 layer in the feature maps mostly match with the key anatomical structures, and thus corroborate the effectiveness of the J-CNN model. The right column of Fig. 4 demonstrates the sequential prediction results by the LSTM model over the video frames. In the detection of all FASP, FFASP, and FFVSP in Fig. 4, the prediction curves share a good consistency with the corresponding ground truths.

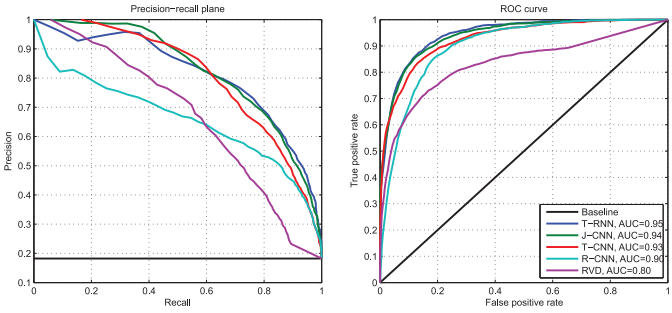

Fig. 5. PR plane and ROC curves of different methods on FASP detection.

C. Comparison of Quantitative Performance

To quantitatively illustrate the efficacy of the proposed T-RNN framework, two most relevant approaches [12], [38] are considered here for performance comparison. The first baseline method is a feature-based approach that exploited the geometric relation and the dedicated features for the task of FASP detection. Specifically, a radial component model and vessel probability map was developed in [12], denoted as RVD, to model the anatomical prior and the geometrical relationship of structures for the plane identification. The second baseline method proposed in [38] is the most related work to this paper. The work of [38] attempted to leverage the transferred knowledge from natural image domains on the detection of FASP in 2-D US videos and this method is called as T-CNN for short, whereas the neural network trained with random initialization is denoted as R-CNN. To further illustrate the effectiveness of the LSTM model on the three detection tasks, we also report the detection performance that is attained solely with J-CNN. The performance report of standard plane detection with only J-CNN model can also help to elucidate that the effect of knowledge-transferring by the multitask framework can mostly yield better boosting of performance than the knowledge learned from natural images. In this paper, we employ four assessment metrics [58] including recall: $R = N_{tp}/(N_{tp} + N_{fn})$, precision: $P = N_{tp}/(N_{tp} + N_{fp})$, F_1 score: $F_1 = 2RP/(R + P)$, and accuracy: $A = (N_{tp} + N_{tn})/(N_{tp} + N_{tn} + N_{fp} + N_{fn})$, where N_{tp}, N_{tn}, N_{fp}, and N_{fn} represents the number of true positives, true negatives, false positives, and false negatives, respectively.

Two comparison schemes are implemented with the basic units of US images and videos. The image-based comparison scheme aims to illustrate the capability of different methods on the differentiation of standard and nonstandard planes over all participating testing images. The second video-based comparison scheme is to see whether the detection algorithms can effectively retrieve the standard plane from an acquired US video. Since the clinical demand for the subsequent biometric measurements and disease diagnosis is to identify the specific standard plane from the scanned US video, the video-based comparison scheme may help to illustrate the clinical applicability of each detection algorithm.

1) Image-Based Evaluation: The image-based comparison results with the four assessment metrics over all comparing methods are shown in Table III. Specifically, the deep learning-based methods of T-RNN, J-CNN, T-CNN, and R-CNN achieve better detection results than the method [12] does on the FASP detection. This may suggest that the engineering of task-specific features may sometimes turn out to be not as useful as the features automatically underlaid by deep learning models. Meanwhile, it can also be observed from Table III that J-CNN and T-CNN [38] outperforms the R-CNN [38] in most assessment metrics. Accordingly, the efficacy of knowledge-transferring on the issues of over-fitting and limited data can be properly substantiated. Furthermore, in most assessment metrics, the J-CNN attains better performance than T-CNN does. This may suggest that the knowledge shared by the three tasks can provide more effective model initialization and learning, as the image domain is the same and the data of the three tasks are relatively relevant (though still quite different). The improvement by the knowledge derived from natural images from ImageNet [59] is relatively limited, probably because the underlying domain gap may be too large to boost the detection performance significantly.

Compared with other methods, our T-RNN method achieves the best performance for the detection of three standard planes. Particularly, for the FASP detection task, a significant outperformance can be observed in Table III, and hence, further suggests the effectiveness of our composite neural network

TABLE IV
RESULTS OF STANDARD PLANE DETECTION ON US VIDEOS

Method	FASP				FFASP				FFVSP			
	A	P	R	$F1$	A	P	R	$F1$	A	P	R	F_1
T-RNN	**0.941**	**0.945**	**0.995**	**0.969**	**0.717**	**0.737**	**0.955**	**0.832**	**0.846**	**0.898**	0.936	**0.917**
J-CNN	0.918	0.922	**0.995**	0.957	0.667	0.736	0.867	0.796	0.769	0.800	**0.952**	0.870
T-CNN [38]	0.904	0.908	**0.995**	0.950	0.633	0.685	0.881	0.771	0.750	0.813	0.907	0.857
R-CNN [38]	0.822	0.826	0.994	0.902	0.567	0.607	0.895	0.723	0.712	0.822	0.841	0.832
RVD [12]	0.762	0.823	0.913	0.865	-	-	-	-	-	-	-	-

framework with the exploration of in- and between-plane cues from US videos. To give more quantitative comparison, the precision-recall plane and receiver operating characteristic curves of all methods on FASP detection task are shown in Fig. 5. The scores of area under the curve obtained by the method of T-RNN, J-CNN, T-CNN, R-CNN, and RVD were 0.95, 0.94, 0.93, 0.90, and 0.80, respectively, further support the outperformance of the proposed T-RNN method.

2) Video-Based Evaluation: To quantitatively assess the capability of standard plane detection from US video by the comparing methods, we follow the same evaluation protocols in [12] for the definition the true positives and true negatives. Specifically, each video is regarded as one testing sample. A true positive identification is defined as the case that a correct standard plane can be successfully detected from a video which encloses at least one standard plane. The true negative case will be confirmed when no standard plane is detected from a video that contains no standard planes. For the methods of T-RNN, J-CNN, T-CNN, and R-CNN, a US video is regarded to have a standard plane if the highest computed score of all member frames is larger than the defined threshold.

The quantitative results of the video-based scheme with respect to the four assessment metrics are reported in Table IV. It can be found that the proposed composite neural network model of T-RNN outperforms other methods in most assessment metrics for the three detection tasks. Specifically, the attained F1 scores are 0.969, 0.832, and 0.917 for the detection of FASP, FFASP, and FFVSP, respectively, whereas the corresponding recall values are all larger than 0.9. Therefore, it can be suggested that most of standard planes can be effectively identified by the T-RNN method from the US videos. Accordingly, the potential applicability of the proposed method to meet the clinical demand can be bright.

The detection system was implemented with the mixed programming technology of Python and C++ based on the open source tool Caffe [36]. It took about 15 h to train the T-RNN model once for all. During the testing, the T-RNN method generally took less than 1 min to identify the standard planes from a video with 40 frames on a workstation equipped with a 2.50 GHz Intel X-eon E5-2609 CPU and an NVIDIA Titan GPU.

IV. DISCUSSION

In this paper, we proposed a composite neural network framework that can effectively discover and fuse the in- and between-plane features to identify desirable standard planes in the US videos. The experimental results corroborate the effectiveness of the usage of multitask framework and the between-plane contextual relation on the detection problems of the FASP, FFASP, and FFVSP. Specifically, by comparing the performances between J-CNN and T-CNN in Table III, the J-CNN model can achieve better performance on the detection of three types of standard planes with the evaluation of all four assessment metrics. Similarly, the J-CNN can mostly achieve better performance as well as in the video-based comparison scheme, see Table IV. It is worth noting that the J-CNN here is co-trained with 900 US videos (37 376 US images), which is significantly less than the millions of natural images in the ImageNet dataset. Although the margin is not large, the outperformance of J-CNN suggests that the multitask framework can leverage the knowledge of thousands of US images as a more effective CNN model initialization than the cross-domain transferring learning does from millions of natural images.

Since the multitask learning is to explore sharable features across different tasks for better generalization, it could help those tasks which are slightly under sampled. However, learning from extremely imbalanced data remains a challenge for most learning techniques. For those tasks with less samples, the sharable features may need to be augmented with the task-specific features to achieve better classification/regression performance. For examples, in the context of semantic characterization of pulmonary nodules, the studies explored the sharable features [60], [61] across different tasks and task-specific features to address the data imbalance issues for the different semantic characteristics of lung nodules in the annotations. With such exploration scheme, the prediction performance can be improved. We happen to have balanced training data for the three tasks in this paper, and hence, the data imbalance problem may not affect our learning scheme seriously. Since the data imbalance issue is a difficult problem in many machine learning contexts, we will explore this issue in the future study.

Referring to the performance comparison between the T-RNN and J-CNN in Tables III and IV, the T-RNN averagely achieves higher scores in both image- and video-based schemes with perceivable margins. It thus can prove that the contextual knowledge learned by the LSTM model can effectively boost the detection performance over all three tasks.

In Table III, the accuracy scores are significantly higher than their corresponding precision and recall scores with respect to all algorithms. Referring to the equations of measurements, it can be found that the computation of accuracy score includes the number of true negatives in the numerator, whereas the calculation of precision and recall scores does not. Since the number of true negative images, i.e., the nonstandard planes, is significantly larger than the number of the true positive

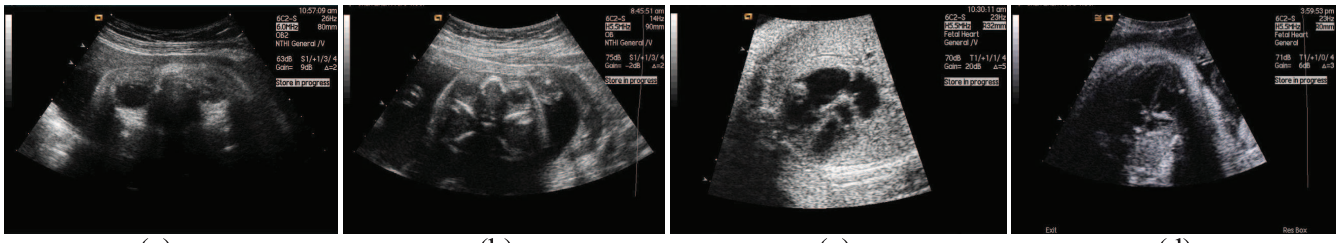

Fig. 6. Examples of false detection results. (a) False positive of FFASP. (b) False negative of FFASP. (c) False positive of FFVSP. (d) False negative of FFVSP.

images (standard planes), the accuracy scores are expected to be larger than the scores of precision and recall. Therefore, the assessment metrics of precision, recall, and $F1$ can be more referential for the evaluation of all comparing algorithms.

Fig. 6 lists some examples of false positives and false negatives by the T-RNN in the FFASP and FFVSP tasks to illustrate the difficulty of these two tasks. The false-positively detected planes may be similar to the standard planes but fail to depict some key structures of each task clearly, e.g., the ocular regions in Fig. 6(a) and ventricular valves in Fig. 6(c). The false negative detections may be due to the confusion with the weak reconstructed acoustic signals, e.g., the left cardiac walls in Fig. 6(d) or the presence of other structures, e.g., the bright structures below the eyes and nose in Fig. 6(b). Generally speaking, the tasks of FFASP and FFVSP are relatively hard as the head and chest regions contain more bone structures and hence the shadowing effect will be more frequently occur.

Although the efficacy of the proposed composite neural network has been well demonstrated in this paper, the developed T-RNN model still has several limitations to be addressed in the future studies. First, the current shape of the T-RNN model may still fall short of satisfying the goal of real-time application. The T-RNN generally takes less than 1 min to identify the standard plane when processing a US video with 40 images. In other words, the time to process one frame takes around 1–2 s, and hence, the operator may easily feel the computational lag with such a processing speed. As a consequence, it is probably not able to generate real-time feedback in the clinical US examination. The computational bottleneck of the T-RNN model lies in the sliding window scanning of the J-CNN model. One potential solution to address the high computational cost of the sliding window scheme may be the replacement of the fully connected layers with the fully convolutional layers [62].

Instead of convolving the image with a small window, the fully convolutional network operates on the whole image with the result in the form of probability map [62], [63]. In this way, the detection process can be possibly sped up as only one pass of forward propagation is carried out and the exhaustive scanning can be prevented. Furthermore, the computation for the standard plane detection may also be accelerated to meet the real-time constraint with the substantial code optimization and parallelization. In this paper, we mainly focus on the algorithm design as well as evaluate the efficacy of the proposed method. We leave the acceleration issue for future studies. The second limitation of the proposed T-RNN model consists in that the current data were acquired from healthy babies and mothers. The generalization to the pathological cases remains unknown. To see the capability of the T-RNN model on the identification of standard planes with abnormalities, we shall continue to collect more clinical data with further IRB approvals.

V. Conclusion

The proposed composite neural network model, i.e., T-RNN, aims to address four major challenges, i.e., high intraclass variation, low interclass variation, limited data, diversity of standard planes, for the computerized detection of fetus standard planes. The T-RNN is able to address the three detection tasks of FASP, FFASP, and FFVSP with the same architecture. With this advantage, the effort to specifically design the detection model for each type of standard plane can be alleviated. The multitask learning framework is introduced here to exploit the shared knowledge across different tasks for reliable model learning and leverage the usage of limited data we have. Meanwhile, with the integration of in- and between-plane cues, the high intraclass and low interclass variation can be further tackled to achieve the current detection performance. The computerized detection of US fetus standard planes is a relatively new topic but crucial to boost the clinical practice. Most previous methods were specifically devised on one dedicated type of standard plane and neglected the contextual cues. This paper proposes a new composite framework with better task generalization and higher identification capability with the fusing of automatically discovered in- and between-plane cues. Accordingly, this paper would shed a light on the potential applicability of composite neural network models on the processing of difficult US image data. Meanwhile, it may be referential to the future studies on the generalization of other US fetus standard planes, and even the plane selection problems of other organs for the US adult examination.

References

[1] J. K. Spencer and R. S. Adler, "Utility of portable ultrasound in a community in ghana," *J. Ultrasound Med.*, vol. 27, no. 12, pp. 1735–1743, 2008.

[2] L. M. Gangarosa, "The practice of ultrasound: A step-by-step guide to abdominal scanning," *Gastroenterology*, vol. 129, no. 4, p. 1357, 2005.

[3] J. Bamber, N. Miller, and M. Tristam, "Diagnostic ultrasound," in *Webb's Physics of Medical Imaging*. Boca Raton, FL, USA: CRC Press, 2012, p. 351.

[4] L. J. Salomon *et al.*, "Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester," *Ultrasound Obstetrics Gynecol.*, vol. 27, no. 1, pp. 34–40, 2006.

[5] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, "Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree," *IEEE Trans. Med. Imag.*, vol. 27, no. 9, pp. 1342–1355, Sep. 2008.

[6] N. J. Dudley and E. Chapman, "The importance of quality management in fetal measurement," *Ultrasound Obstetrics Gynecol.*, vol. 19, no. 2, pp. 190–196, 2002.

[7] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, and S. K. Zhou, "Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Athens, Greece, 2016, pp. 487–495.

[8] L. Zhang, S. Chen, C. T. Chin, T. Wang, and S. Li, "Intelligent scanning: Automated standard plane selection and biometric measurement of early gestational sac in routine ultrasound examination," *Med. Phys.*, vol. 39, no. 8, pp. 5015–5027, 2012.

[9] L. Wu *et al.*, "FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2017.2671898.

[10] Amer. Inst. Ultrasound Med., "AIUM practice guideline for the performance of obstetric ultrasound examinations," *J. Ultrasound Med. Official J. Amer. Inst. Ultrasound Med.*, vol. 29, no. 1, pp. 157–166, 2010.

[11] R. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward, "Localizing target structures in ultrasound video—A phantom study," *Med. Image Anal.*, vol. 17, no. 7, pp. 712–722, 2013.

[12] D. Ni *et al.*, "Standard plane localization in ultrasound by radial component model and selective search," *Ultrasound Med. Biol.*, vol. 40, no. 11, pp. 2728–2742, 2014.

[13] B. Rahmatullah, I. Sarris, A. Papageorghiou, and J. A. Noble, "Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using AdaBoost," in *Proc. IEEE Int. Symp. Biomed. Imag. Nano Macro*, Chicago, IL, USA, 2011, pp. 6–9.

[14] A. Abuhamad, P. Falkensammer, F. Reichartseder, and Y. Zhao, "Automated retrieval of standard diagnostic fetal cardiac ultrasound planes in the second trimester of pregnancy: A prospective evaluation of software," *Ultrasound Obstetrics Gynecol.*, vol. 31, no. 1, pp. 30–36, 2008.

[15] B. Rahmatullah, A. T. Papageorghiou, and J. A. Noble, "Integration of local and global features for anatomical object detection in ultrasound," in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, Nice, France, 2012, pp. 402–409.

[16] M. Sofka, J. Zhang, S. Good, S. K. Zhou, and D. Comaniciu, "Automatic detection and measurement of structures in fetal head ultrasound volumes using sequential estimation and integrated detection network (IDN)," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1054–1070, May 2014.

[17] D. Ni *et al.*, "Selective search and sequential detection for standard plane localization in ultrasound," in *Abdominal Imaging. Computation and Clinical Applications*. Heidelberg, Germany: Springer, 2013, pp. 203–211.

[18] H. Chen, D. Ni, X. Yang, S. Li, and P. A. Heng, "Fetal abdominal standard plane localization through representation learning with knowledge transfer," in *Machine Learning in Medical Imaging*. Heidelberg, Germany: Springer, 2014, pp. 125–132.

[19] M. A. Maraci, R. Napolitano, A. Papageorghiou, and J. A. Noble, "Searching for structures of interest in an ultrasound video sequence," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2014, pp. 133–140.

[20] B. R. Benacerraf, "Three-dimensional fetal sonography: Use and misuse," *J. Ultrasound Med.*, vol. 21, no. 10, pp. 1063–1067, 2002.

[21] J. Shi *et al.*, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, Jun. 2016.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[23] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[25] H. Qiao, Y. Li, F. Li, X. Xi, and W. Wu, "Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2335–2347, Oct. 2016.

[26] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 221–248, 2017.

[27] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[28] J. Shi, J. Wu, Y. Li, Q. Zhang, and S. Ying, "Histopathological image classification with color pattern random binary hashing based PCANet and matrix-form classifier," *IEEE J. Biomed. Health Inform.*, to be published, doi: 10.1109/JBHI.2016.2602823.

[29] H. Chen *et al.*, "3D fully convolutional networks for intervertebral disc localization and segmentation," in *Proc. Int. Conf. Med. Imag. Virtual Reality*, Bern, Switzerland, 2016, pp. 375–382.

[30] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[31] Q. Dou *et al.*, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Athens, Greece, 2016, pp. 149–157.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, 2015, pp. 234–241.

[33] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, 2014, pp. 512–519.

[36] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 675–678.

[37] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, Lille, France, 2015, pp. 97–105.

[38] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[39] H. Chen *et al.*, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 507–514.

[40] L. Medsker and L. Jain, "Recurrent neural networks," in *Design and Applications*. Washington, DC, USA: CRC Press, 2001.

[41] A. Graves *et al.*, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[42] Z. Yi, J. C. Lv, and L. Zhang, "Output convergence analysis for a class of delayed recurrent neural networks with time-varying inputs," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 1, pp. 87–95, Feb. 2006.

[43] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385. Berlin, Germany: Springer, 2012.

[44] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 2625–2634.

[45] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3128–3137.

[46] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 3104–3112.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*. New York, NY, USA: IEEE Press, 2001.

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[50] A. Gupta, M. S. Ayhan, and A. S. Maida, "Natural image bases to represent neuroimaging data," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, 2013, pp. 987–994.

[51] H. Chen *et al.*, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, pp. 135–146, Feb. 2017.

[52] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proc. 14th Int. Conf. Artif. Intell. Stat. JMLR W CP Volume*, vol. 15. Lille, France, 2011, pp. 315–323.
[53] S. Wager, S. Wang, and P. S. Liang, "Dropout training as adaptive regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 351–359.
[54] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, pp. 1–43, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850
[55] W. Zaremba and I. Sutskever, "Learning to execute," *CoRR*, vol. abs/1410.4615, pp. 1–25, 2014. [Online]. Available: http://arxiv.org/abs/1410.4615
[56] R. J. Williams and D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," in *Back-Propagation: Theory, Architectures and Applications*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1995, pp. 433–486.
[57] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 818–833.
[58] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval*. Heidelberg, Germany: Springer, 2005, pp. 345–359.
[59] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[60] S. Chen *et al.*, "Bridging computational features toward multiple semantic features with multi-task regression: A study of CT pulmonary nodules," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Cham, Switzerland, 2016, pp. 53–60.
[61] S. Chen *et al.*, "Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 802–814, Mar. 2017.
[62] H. Chen, Q. Dou, X. Wang, J. Qin, and P.-A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 1160–1166.
[63] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3431–3440.

Hao Chen (S'14) received the B.E. degree in information engineering from Beihang University, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

His current research interests include medical image analysis, deep learning, and health informatics.

Mr. Chen was a recipient of the Hong Kong Ph.D. Fellowship.

Lingyun Wu received the B.S. degree in biomedical engineering from the South Central University for Nationalities, Wuhan, China, in 2014. She is currently pursuing the master's degree with the School of Biomedical Engineering, Shenzhen University, Guangdong, China.

Her current research interests include intelligent ultrasound diagnosis and pattern recognition.

Qi Dou (S'14) received the B.E. degree in biomedical engineering from Beihang University, Beijing, China, in 2014. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

Her current research interests include medical image analysis, deep learning, computer-aided detection, and segmentation.

Jing Qin (M'16) received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2009.

He is an Assistant Professor with the School of Nursing, Hong Kong Polytechnic University, Hong Kong, where he is also a Key Member with the Centre for Smart Health. He has participated in over 10 research projects and published over 90 papers in major journals and conferences in the below areas. His current research interests include virtual/augmented reality for healthcare and medicine training, medical image processing, deep learning, visualization and human-computer interaction, and health informatics.

Shengli Li received the master's degree in radiology from the Xiang Ya School of Medicine, Hunan, China, in 1994.

He is currently a Chief Physician and a Professor with the Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital, Nanfang Medical University, Guangdong, China. His current research interest includes ultrasound diagnosis.

Jie-Zhi Cheng (M'16) received the Ph.D. degree in biomedical engineering from National Taiwan University, Taipei, Taiwan, in 2013.

He is currently an Associate Professor with the School of Biomedical Engineering, Shenzhen University, Guangdong, China. His current research interests include medical image analysis, computer-aided diagnosis and intervention, pattern recognition, and machine learning.

Dong Ni received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2009.

He is currently an Associate Professor with the School of Biomedical Engineering, Shenzhen University, Guangdong, China. His current research interests include ultrasound image analysis, image guided surgery, and pattern recognition.

Pheng-Ann Heng (M'92–SM'06) received the Ph.D. degree in computer science from Indiana University, Indianapolis, IN, USA.

He is currently a Professor with the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong, where he is also the Director of the Virtual Reality, Visualization, and Imaging Research Centre. He is also the Director of the Research Center for Human-Computer Interaction, Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include virtual reality applications in medicine, visualization, medical imaging, human-computer interfaces, rendering and modeling, interactive graphics, and animation.