Learning to Segment Brain Anatomy from 2D Ultrasound with Less Data

V Jeya Maria Jose, *Student Member, IEEE,* Rajeev Yasarla, *Student Member, IEEE,*, Puyang Wang, *Student Member, IEEE,*, Ilker Hacihaliloglu, *Member, IEEE,* and Vishal M. Patel, *Senior Member , IEEE*

***Abstract*—Automatic segmentation of anatomical landmarks from ultrasound (US) plays an important role in the management of preterm neonates with a very low birth weight due to the increased risk of developing intraventricular hemorrhage (IVH) or other complications. One major problem in developing an automatic segmentation method for this task is the limited availability of annotated data. To tackle this issue, we propose a novel image synthesis method using multi-scale self attention generator to synthesize US images from various segmentation masks. We show that our method can synthesize high-quality US images for every manipulated segmentation label with qualitative and quantitative improvements over the recent state-of-the-art synthesis methods. Furthermore, for the segmentation task, we propose a novel method, called Confidence-guided Brain Anatomy Segmentation (CBAS) network, where segmentation and corresponding confidence maps are estimated at different scales. In addition, we introduce a technique which guides CBAS to learn the weights based on the confidence measure about the estimate. Extensive experiments demonstrate that the proposed method for both synthesis and segmentation tasks achieve significant improvements over the recent state-of-the-art methods. In particular, we show that the new synthesis framework can be used to generate realistic US images which can be used to improve the performance of a segmentation algorithm.**

***Index Terms*—Ultrasound, brain, deep learning, ventricle, septum pellecudi, preterm neonate, confidence map, segmentation, synthesis.**

I. Introduction

According to the World Health Organization, 15 million babies are born preterm each year [1]. Although, advancements made in neonatal care have increased the survival rates, majority of these infants are at risk for long-term complications such as cerebral palsy, cognitive-behavioral and learning impairments. In premature infants, one of the most common brain injury is intraventricular hemorrhage (IVH) [2]. These hemorrhages result in ventricle dilation, which can lead to serious brain damage if not properly treated. Ventricle dilation is also associated with white matter atrophy (hydrocephalus ex-vacuo). Therefore, monitoring of ventricle volume change in neonates is clinically important in order to determine the correct intervention. On the other hand absence of septum pellucidum is used as a valuable landmark for the diagnosis of abnormalities, such as septo-optic dysplasia, in the central nervous system (CNS) [3], [4]. The main imaging modality currently employed for monitoring brain abnormalities in preterm neonates is two-dimensional (2D) ultrasound (US) due to its real-time safe imaging capabilities. However, high levels of noise and various imaging artifacts, and irregular shape deformation of ventricles, results in the inability to localize the site and extent of brain injury, or to predict neurologic outcomes in identifying IVH or other abnormalities from US data. Being a user dependent imaging modality causes additional difficulties during data collection since a single-degree deviation angle by the operator can reduce the signal strength by 50%. Current clinical practice involves manual measurement of ventricle or investigation of septum pellucidum presence from the collected scans by clinicians. Due to previously mentioned difficulties, related to US imaging, this is an error prone and time consuming process.

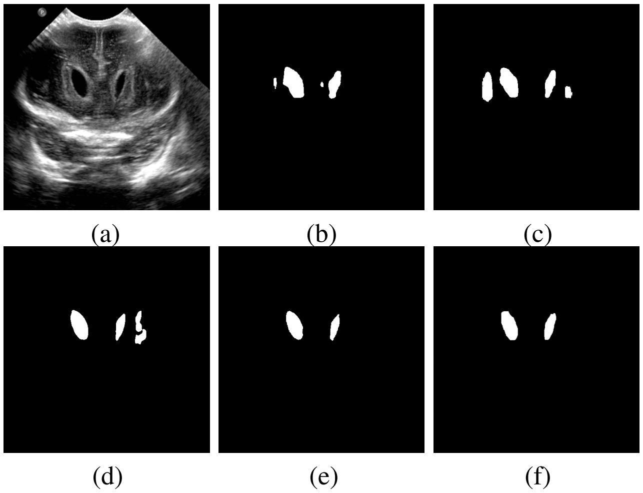

Fig. 1. (a) Original brain US image. Brain ventricular segmentation obtained using (b) pix2pix [5], (c) U-net [6], (d) Wang et al.[7], (e) CBAS (ours). (f) ground-truth brain ventricular regions.

In order to automate the ventricle segmentation and measurement process, various groups have proposed automatic segmentation methods. In [9], a fully automated atlas-based

Jeya Maria Jose V., is with the Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, e-mail: jvalana1@jhu.edu

Rajeev Yasarla, is with the Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, e-mail: ryasarl1@jhu.edu

Puyang Wang, is with the Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608, e-mail: pwang47@jhu.edu

Ilker Hacihaliloglu, is with the Department of Biomedical Engineering, Rutgers, The State University of New Jersey, e-mail: ilker.hac@soe.rutgers.edu

Vishal M. Patel, is with the Whiting School of Engineering, Johns Hopkins University, e-mail: vpatel36@jhu.edu

Manuscript received...

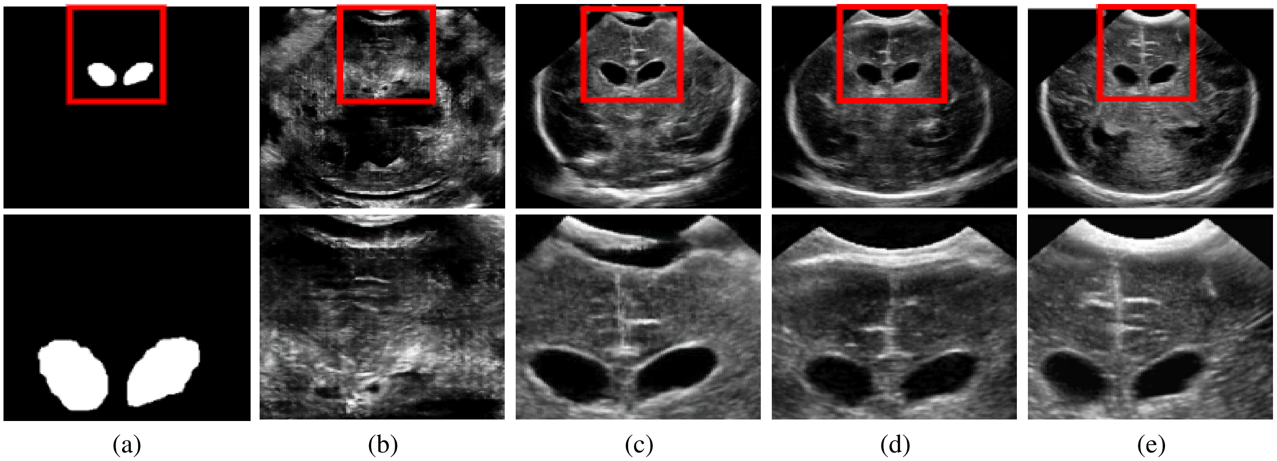

Fig. 2. (a) Input segmentation mask. Synthesized image using (b) pix2pix [5] (c) pix2pixHD [8] (d) MSSA (ours) (e) Original image corresponding to the segmentation mask in (a). The second row consists of the zoomed portions of the image inside the red box in the first row.

segmentation pipeline was developed for segmenting 3D volumetric US data. Validation results performed on 30 3D US scans, obtained from 14 patients, achieved a mean Dice similarity coefficient (DSC) and maximum absolute distance of 76.5$\%$ and 1 mm, respectively. The reported computation time for segmenting a single 3D volume was 54 mins [9]. Atlas-based volumetric US segmentation was also proposed in [10]. Validation performed on 16 subject scans achieved a mean DSC of 0.70. Computation time was not reported. A semi-automatic method, for segmenting volumetric US scans, was proposed in [11]. Mean absolute distance between the manual and semi-automatically segmented contours was 2.17 mm. Subject size and processing time was not reported [11]. In order to improve the accuracy and computation time, methods based on deep learning have been investigated [12], [7]. In [12], a U-net based [6] network architecture was proposed. Reported mean DSC value and computation time were 0.81 and 5 seconds per volume (0.01 seconds per slice) respectively for 15 volumes obtained from 14 patients. In [7], a multi-scale-based network architecture was proposed for segmentation of 2D US scans. Validation studies performed on 687 scans, obtained from 10 subjects, achieved a mean DSC value of 0.90 with a computation time of 0.02 seconds.

Although, deep learning methods have resulted in increased accuracy and computation time, most of the previous work has been validated on scans with enlarged ventricles. If the foreground anatomical structure is to be segmented, traditional convolutional neural network (CNN) architectures fail since there is not enough positional information to localize small brain anatomy as they are significantly smaller compared to the background anatomical structure. The same is also valid for segmenting densely packed small brain anatomy (small ventricles and septum pellecudi appearing in the middle of the US scan). Finding small anatomical structure using a CNN architecture is difficult since resolution of small features is gradually lost and resulting coarse features can miss the details of small structures [13]. For example, methods like pix2pix [5], U-net [6], and Wang et al. [7] fail to segment the brain ventricular region from the US images as shown in Fig. 1. These methods end up segmenting the non-ventricular region as the brain ventricular region. This is mainly due to the lack of special attention given to small ventricles while learning the network weights. Finally, due to the high complexity and variability in the ventricles shape, the traditional CNN architectures result in over or under segmentation (Fig. 1).

To address this problem, we propose a method called, Confidence-guided Brain Anatomy Segmentation (CBAS) network, where we make use of the aleatoric uncertainty and define confidence scores at each pixel which are data dependent. Uncertainty can be modeled in two ways – epistemic and aleatoric uncertainties as explained in [14], [15]. In order to achieve better performance in tasks like medical image segmentation, [16], [17], [18] modeled epistemic uncertainty for learning the CNN network weights. Uncertainty has also been used to efficiently leverage unlabeled data in a semi-supervised setting for atrium segmentation [19]. Some methods like [20], [21], [22] use boundary information to enhance medical image segmentation. To handle different brain anatomy structures in 2D US scans, we define data dependent aleatoric uncertainty as the confidence scores that are computed by the confidence blocks in CBAS. These blocks essentially indicate how confident the CBAS network is about the segmentation output. This confidence score will be low for the regions where the error is high and vice-versa. Thus CBAS learns to differentiate the erroneous regions and gives special attention to those regions in subsequent layers while computing the segmentation output. We present a novel method for fully automatic ventricles and septum pellecudi segmentation with varying size from 2D US scans. Note that this is the first approach that uses uncertainty in 2D US segmentation. We validate our method on 1629 US scans obtained from 20 different subjects.

One of the major problems in medical image analysis is the limited number of annotated data. Obtaining clinical annotations is also a difficult, expensive and time consuming process as expert radiologists are needed. For very specific tasks like the one addressed in this paper, the availability of datasets is also very scarce. As most of the current state-of-the-art segmentation methods require a considerable amount of data to train the network, using them for tasks with less data does not guarantee a good performance. As a result, novel image synthesis methods are proposed in the literature to synthesize meaningful high quality data that could be added to the training dataset.

Over the past few years, image synthesis and image-to-image translation tasks have been dominated by Generative Adversarial Networks (GANs) [23] and its variations. In this approach, a generator is trained to synthesize an image from random noise while a discriminator, which is trained on both real and synthesized images tries to classify whether the image is real or was synthesized by the generator (i.e. fake). Both networks are trained in a min-max way such that they act as adversaries of each other. While using GANs in medical imaging to synthesize new images solves the issue of limited availability of data, the problem of annotations still exists in this setup. Isola et al. [5] proposed using a conditional generative adversarial network (cGAN) [24] to solve the image-to-image translation tasks where the network is trained to learn the mapping between an image across two different domains. In the medical imaging community, several works ([25], [26], [27], [28], [29], [30]) have adapted this idea to synthesize images from one modality to another modality such as MRI to CT, T1 MRI to T2 MRI etc. Since this method can be used for any translation task, it can be used for image synthesis from segmentation labels where the network is trained to translate the segmentation mask of an image into a realistic US image. Zhao et al. [31] showed that multiple realistic-looking retinal images can be synthesized from just the annotation masks using this method. Bailo et al. [32] used cGAN to generate blood smear image data from segmentation masks corresponding to microscopic images. Diverse set of new images were also achieved by manipulating the segmentation labels. Jaiswal et al. [33] used a capsule cGAN to synthesize microscopic data of cortical axons.

Though many methods exist for medical image synthesis, most of them only deal with generating low resolution images. Different from MRI and CT image synthesis using GAN, ultrasound image synthesis is a more difficult task. One of the main reason is the manual data collection aspect of ultrasound. Even with healthy pathology, the orientation of the ultrasound transducer will effect the quality of the collected data and subsequent image analysis methods. In Fujioka et al. [34] breast US images were synthesized using a GAN-based approach [35]. The authors however fail to show any quantitative analysis or usefulness of the synthesized images. In [36], fetal US images were synthesized from tracked B-mode US data. Validation experiments were performed on US data obtained from a fetus ultrasound examination phantom. Simulation of realistic in vivo US data is a more challenging problem as soft tissue properties vary significantly depending on the imaged subject and orientation of the transducer [36]. In [37], GANs were used to simulate intravascular US (IVUS) data using convolution networks. Most of the synthesis methods that use only convolutional networks fail to capture very long range dependencies in the image due to the relatively low receptive field of convolution. This can be clearly seen by comparing the performance of different synthesis methods as shown in Fig. 2. It can be observed that pix2pix [5] synthesizes very poor quality image and pix2pixHD [8] fails to capture the fine details of the ultrasound image towards the edges. To tackle this issue, we propose a novel attention-based method that can synthesize realistic brain US images from a ventricle and septum pellecudi segmentation masks. We use a multi-scale generator architecture with multi-scale self-attention modules that guides the network to capture the long range dependencies while also synthesizing high resolution images. A sample synthesized image using our method is shown in Fig 2(d). As compared to the other synthesis methods, the proposed method produces sharper images from the input segmentation masks. Using the proposed synthesis model, numerous realistic US images can be synthesized by manipulating the segmentation masks that is fed into the network. As the images are directly synthesized from the manipulated segmentation masks, there is no need for annotation of the synthesized data. By performing extensive experiments, we show that these synthesized images, when added to the training data, increase the performance of the segmentation network.

This paper makes the following contributions:

- A novel synthesis network is proposed using a multi-scale generator guided by self-attention modules to synthesize realistic US images from the segmentation masks.
- A novel US image segmentation method, called CBAS, is proposed which generates the segmentation maps at different scales along with the confidence maps, to guide subsequent layers the network by blocking the propagation of errors in the segmentation map at lower scale, while computing final output segmentation.
- A novel loss function is introduced to train CBAS which makes use of the computed confidence maps and the corresponding segmentation maps.
- Extensive experiments are conducted to show the significance of the proposed synthesis and segmentation networks. Furthermore, an ablation study is conducted to demonstrate the effectiveness of different parts of our networks. We also show that the synthesized images are useful as they can be used to improve the segmentation performance.

Rest of the paper is organized as follows. Details of the proposed uncertainty-guided segmentation method are given in Section II. Section III gives details regarding the proposed self-attention based synthesis method. Experimental results as well as ablation study details are given in Section IV. Finally, Section V concludes the paper with a brief summary and discussion.

II. Confidence-guided Brain Anatomy Segmentation (CBAS)

Let the set of brain US scans be denoted as $\mathcal{B}$ and the corresponding set of brain ventricle segmentation maps as $\mathcal{S}$. Our aim is to estimate the brain ventricle segmentation map $\hat{s}$ for a given brain US scan $x \in \mathcal{B}$. To address this problem unlike many deep learning-based methods that directly estimate the brain ventricle segmentation map, we take a different approach in which we first estimate the segmentation map

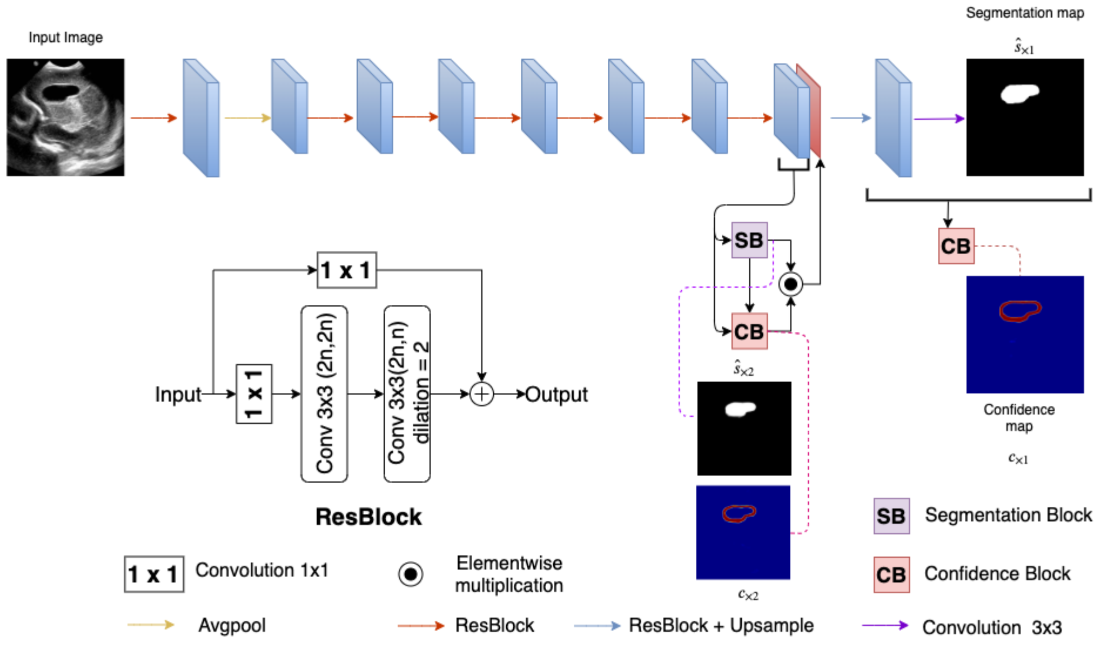

Fig. 3. An overview of the proposed CBAS network. The aim of the CBAS network is to estimate the brain anatomy segmentation for the given brain US image. CBAS learns the segmentation maps and computes confidence maps to guide the network. To achieve this, we introduce SB and CB networks and feed their outputs to the subsequent layers. Note that in the confidence maps blue corresponds to 1 and red corresponds to 0.

s and the corresponding confidence map c. We define the confidence map c, that represents the confidence score at each pixel which resembles the measure of how much the network is certain about the computed value in the segmentation map. Our proposed method, CBAS, judiciously combines the segmentation and confidence information at lower scales to block the propagation of errors in the segmentation $\hat{s}_{\times2}$ while computing the final segmentation map $\hat{s}$. Fig. 3 gives an overview of the proposed CBAS network. As it can be seen from this figure, we estimate the segmentation map $\hat{s}_{\times2}$ and the confidence map $c_{\times2}$ at scale $\times2$ (0.5 scale of x) and they are fed back to the subsequent layers in a way that blocks the errors in $\hat{s}_{\times2}$ using $c_{\times2}$.

In CBAS, we estimate the segmentation maps at two different scales, $i \in \{\times1, \times2\}$, i.e $\hat{s}_{\times1}$ (same size as x) and $\hat{s}_{\times2}$ (0.5 scale as x), and the corresponding confidence maps $c_{\times1}$ and $c_{\times2}$. To estimate these segmentation maps, we construct our base network (BN) using U-Net [6] and Res-Net [38] architectures with the ResBlock as our basic building block. To increase the receptive field size, we introduce dilation convolutions in the ResBlock, as shown in Fig. 3, where Conv $l\times\ l\ (m,n)$ contains instance normalization [39], Rectified Linear Unit (ReLU), Conv $(l \times l)$ - convolutional layer with kernel of size $l \times l$, where m and n are the number of input and output channels, respectively. Note that all convolutional layers in BN are densely connected [40]. The BN network consists of the following sequence of layers:
ResBlock(1,32)-Avgpool-ResBlock(32,32)-Avgpool-ResBlock(32,32)-
ResBlock(32,32)-ResBlock(32,32)-ResBlock(32,32)-Upsample-ResBlock(32,32)-
Upsample-ResBlock(33,16)-Conv3 $\times$ 3(16,1),
where Avgpool is the average pooling layer, and Upsample is the upsampling convolution layer.

A. CBAS Network

Segmentation networks are prone to misclassify the labels near the edges of brain ventricles. Hence a brain ventricle segmentation method requires special attention in those regions where the network may go wrong. To address this issue, one can estimate brain ventricle segmentation at different scales, and estimate the confidence map which indicates the regions where the method can go wrong. Confidence map highlights the regions where the network is certain about the segmentation values by producing high confidence values (i.e nearly 1) and assigning low confidence scores for those pixels where the network is uncertain. In this way, highlighting the regions in the confidence map and combining them with the segmentation map, we block the propagation of errors in segmentation, and make the network more attentive in the erroneous regions. To estimate these pairs of segmentation and the corresponding confidence map, we introduce Segmentation Block (SB) and Confidence Block (CB) in our base network (BN) and construct our CBAS network as shown in Fig. 3.

B. Segmentation and Confidence Blocks

Feature maps at half-scale are given as input to the Segmentation Block (SB) to compute the segmentation map $s_{\times2}$. SB has a sequence of four convolutional layers. We feed the estimated segmentation maps and the feature maps as inputs to CB for computing the confidence score at every pixel, which indicates how certain the network is about the segmentation value. CB has a sequence of four convolutional layers. Details

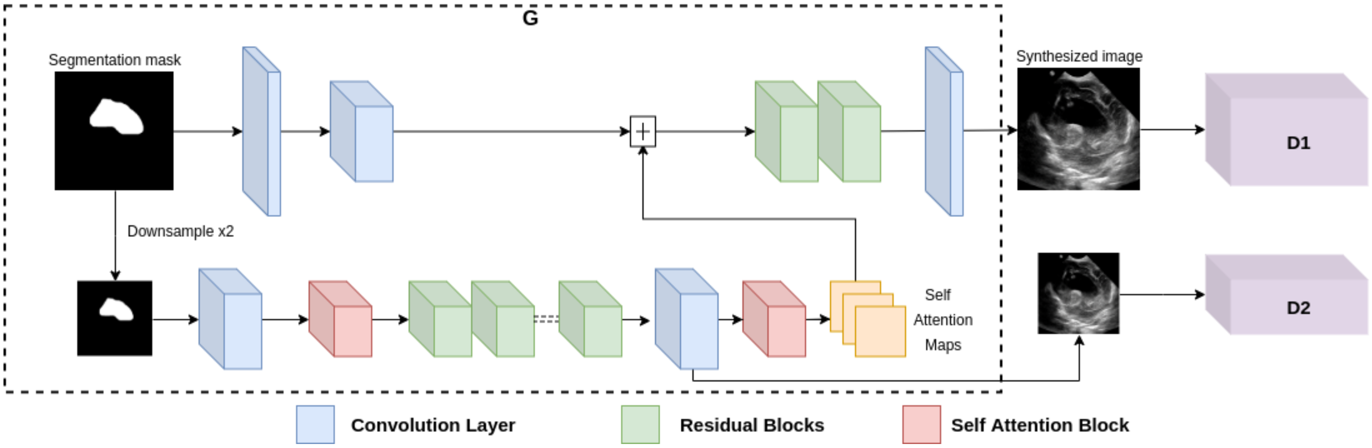

Fig. 4. An overview of the proposed MSSA network. The MSSA network takes in a segmentation mask and synthesizes the corresponding realistic looking synthetic ultrasound image. G denotes the generator. D1 and D2 denote the discriminators across each scale.

of convolutional layers in SB and CB blocks are shown in Table V (in Appendix A).

Given an US image x, we estimate the segmentation maps ($\hat{s}_{\times 1}$ and $\hat{s}_{\times 2}$) as well as the corresponding confidence maps ($c_{\times 1}$ and $c_{\times 2}$) as shown in Fig. 3. We propose a confidence-guided loss function to train the CBAS network which uses a pair of segmentation and confidence maps (i.e $\{\hat{s}_{\times 1},\ c_{\times 1}\}$ and $\{\hat{s}_{\times 2},\ c_{\times 2}\}$).

C. Loss for CBAS

Likelihood-based framework can be used to optimize the CBAS network parameters (θ) as follows,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(f_\theta(x)|x;\theta) = \underset{\theta}{\operatorname{argmax}}\, P(\hat{s}|x;\theta), \tag{1}$$

where $P(.)$ is the probability function, $f_\theta(.)$ represents the CBAS network, $\hat{s} = f_\theta(x)$. To find the optimal network parameters $\hat{\theta}$, $P(\hat{s}|x;\theta)$ needs to be maximized. For simplicity to solve this optimization problem, let us assume $P(\hat{s}|x;\theta)$ is a Gaussian distribution. As our goal is to minimize the error between $\hat{s}$ and the actual segmentation map (s) of x, we denote the mean of distribution $P(.)$ as s and variance as σ^2. Thus our objective from Eq. 1 becomes,

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, \log(P(\hat{s}|x;\theta)) \\ \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} -\frac{1}{\sigma^2}\|\hat{s}-s\|_2^2 + \log(\frac{1}{\sigma^2}).\end{aligned} \tag{2}$$

In the above Eq. 2, variance (σ^2) can be inferred in two ways as explained in [14], [15], (i) Epistemic uncertainty, which is explained as the model uncertainty given enough data to train, and (ii) Aleatoric uncertainty that captures noise inherent in the observations, which is data dependent. Epistemic uncertainty can be formulated as variational inference to compute variance. Aleatoric uncertainty can be formulated as MAP (maximum-aposterior) or ML (maximum-likelihood) inference. Here, in our method we attempt to address uncertainty caused in outputs due to different sizes of brain ventricles and sensor noise which is inherent in brain US images. Following the ML inference, we can formulate $\frac{1}{\sigma^2}$ as the confidence score (c), i.e finding a confidence score at every pixel in the output which depends on the input brain ultrasound scan. We compute these confidence scores using CB (confidence block) as explained in the earlier sections. Computing these confidence scores benefits us in learning the network weights as the erroneous regions have low confidence scores. Note that, to capture the erroneous regions, the confidence score should be estimated pixel-wise. We will benefit in guiding the network by recognizing the regions which are prone to make errors if we estimate the confidence scores pixel-wise. Note that values in the confidence map at every position will be in the range of $[0, 1]$. Thus, we modify Eq. 2, to accommodate these properties of the confidence scores as follows

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}} -\frac{1}{\sigma^2}\|\hat{s}-s\|_2^2 + \log\left(\frac{1}{\sigma^2}\right) \\ \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \sum_j \sum_k -\frac{1}{\sigma_{jk}^2}\|\hat{s}_{jk}-s_{jk}\|_2^2 + \log\left(\frac{1}{\sigma_{jk}^2}\right) \\ \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \sum_j \sum_k -c_{jk}\|\hat{s}_{jk}-s_{jk}\|_2^2 + \lambda\log(c_{jk}),\end{aligned} \tag{3}$$

where j and k are dimensions of s and $\hat{s}$. Since our task is segmentation, we construct our Loss function by replacing the L2-norm in the above Eq. 3 with the cross-entropy loss as follows,

$$\mathcal{L}(\hat{s}, s) = \sum_j \sum_k c_{jk}\mathcal{L}_{CE}(\hat{s}_{jk}, s_{jk}) - \lambda\log(c_{jk}), \tag{4}$$

where

$$\mathcal{L}_{CE}(\hat{s}_{jk}, s_{jk}) = -s_{jk}\log(\hat{s}_{jk}) - (1-s_{jk})\log(1-\hat{s}_{jk}).$$

Since we are estimating the segmentation maps at two different scales, we extend this loss to train CBAS as follows,

$$\mathcal{L}_{final} = \sum_{i\in\{\times 1,\ \times 2\}} \sum_j \sum_k c_{i_{jk}}\mathcal{L}_{CE}(\hat{s}_{i_{jk}}, s_{i_{jk}}) - \lambda\log(c_{i_{jk}}). \tag{5}$$

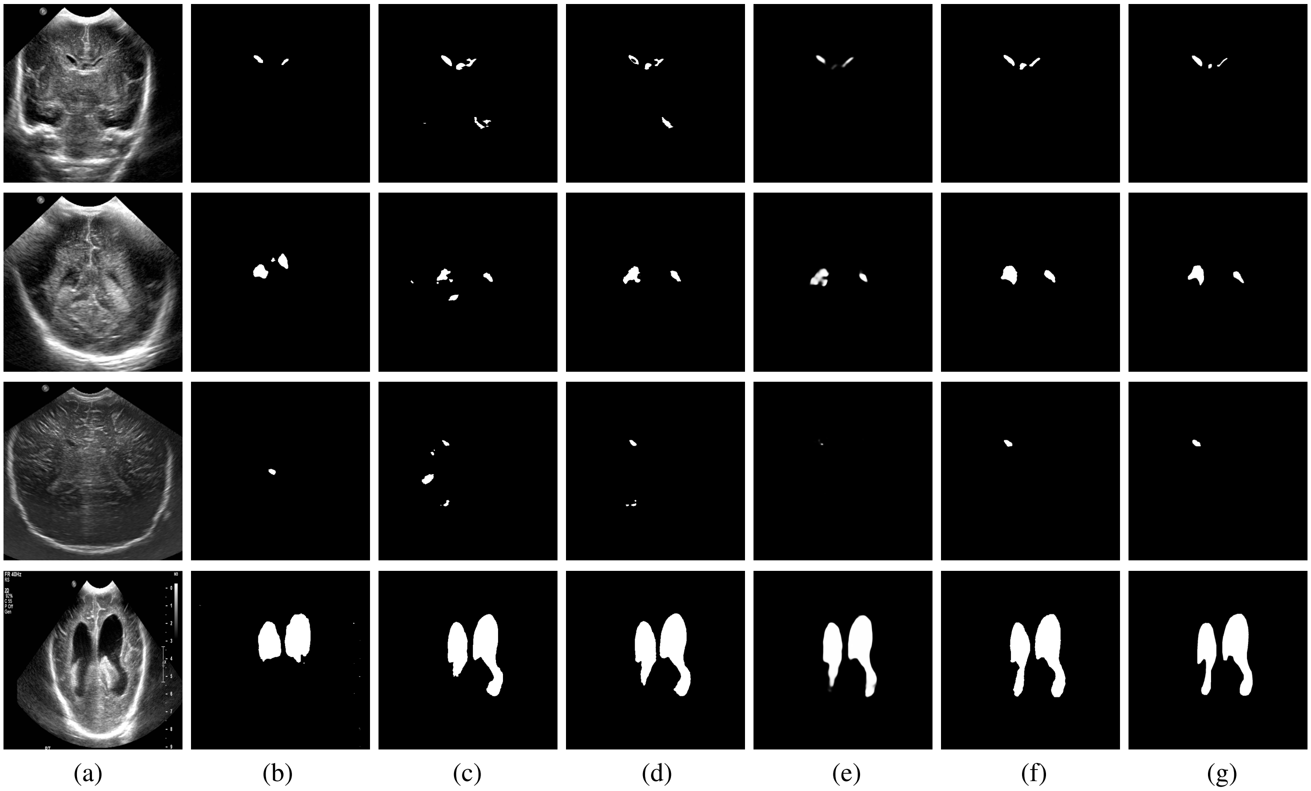

Fig. 5. Qualitative results on test images. (a) Input brain ultrasound image. (b) pix2pix [5]. (c) U-Net [6]. (d) UDe-Net [6], [40]. (e) Wang et.al [7]. (f) CBAS (ours) (g) Ground-truth ventricle segmentation.

TABLE I
COMPARISON WITH PIX2PIX [5], U-NET[6], UDE-NET[6], [40], WANG ET AL.[7]. RESULTS SHOWN CORRESPOND TO MEAN VALUES AND VARIANCE.

Method	DICE	IoU(%)	Parameters
pix2pix[5]	0.8584 ± 0.025	77.96 ± 0.031	11.1MB
U-Net[6]	0.8538 ± 0.024	77.90 ± 0.028	6.7MB
UDe-Net[6], [40]	0.8598 ± 0.017	78.09 ± 0.018	6.7MB
Wang et.al[7]	0.8725 ± 0.016	79.28 ± 0.014	24.9MB
CBAS	0.8813 ± 0.008	80.25 ± 0.010	6.7MB
CBAS (with synthetic data generated using MSSA)	**0.8901 ± 0.063**	**81.03 ± 0.061**	6.7MB

III. MULTI-SCALE SELF ATTENTION (MSSA) GUIDED SYNTHESIS

As the CBAS network is data-driven like most other deep learning methods, the performance of it is based on the size of training dataset. Collection of any medical image data and performing annotations of the same is a cumbersome and expensive process. An approach to deal with this issue is to generate meaningful synthetic data which can be used to boost the segmentation performance. To this end, we propose an image synthesis network that is trained to generate real-looking US images given segmentation masks. Inspired from [8], we propose a multi-scale generator and discriminator networks to produce high-quality US images. Multi-scale networks have been used to generate stable high-resolution images [8]. However, they still fail to capture long-range dependencies in the US images. This makes the synthesized images look unrealistic with many artifacts near the edges of the anatomical structures. To avoid this from happening, we propose a self-attention guided method where the self-attention module [41] is used to leverage the small-range capturing ability of the convolution blocks. The proposed network is called Multi-Scale Self-Attention network (MSSA) Network.

A. MSSA Network

Using the same notations as in Section II, the problem statement can be viewed as an image translation task of synthesizing $\hat{x}$ from a given brain ventricle segmentation mask s. During training, a segmentation map s such that $s \in \mathcal{S}$ is taken as input and its corresponding US scan x such that $x \in \mathcal{B}$ is taken as the ground truth. The network we propose has a multi-scale generator architecture where the first part of the generator operates on the original scale of the segmentation mask s and the second part of the generator operates on a down scaled (by 2) version of the segmentation mask $s_{\times 2}$. The proposed self-attention guided block operates on the down scaled version. Each self attention module [42][41] has three 1×1 convolution filters that are applied to the convolution feature maps. The output of each of the 1×1 convolution layers can be represented as

$$\begin{aligned} K(x) &= W_k x, \\ Q(x) &= W_q x, \\ V(x) &= W_v x, \end{aligned}$$

where W_k,W_q and W_v are the 1×1 convolution filters and x is the convolutional feature maps. To get the self-attention feature maps, we perform dot product as follows

$$\alpha_{i,j} = softmax(K(x_i)^T Q(x_j))$$

$$o_j = \sum_{i=1}^{N} \alpha_{i,j} V(x_i)$$

where $\alpha_{i,j}$ indicates the amount of attention the model gives while synthesizing the j^{th} position from the i^{th} location. The output self attention feature maps is the collection of the individual feature vectors o_j where j goes from 1 to N.

The segmentation mask is first passed through a convolutional layer followed by an attention module which captures the dependencies of the image in its feature space. It is followed by a series of residual blocks [38]. We use another self-attention module at the end of residual blocks to get the self-attention feature maps. These are concatenated with the feature maps that are generated from the generator at the original scale. The resulting concatenated feature maps are then further passed through the residual blocks before passing them through transpose convolution layers to get the US image. Owing to the high resolution of the synthesized image, we use a two scale discriminator that works on the original as well as the down scaled (by 2) version of the real and synthesized image. The discriminator architecture across both scales are patch based fully convolutional networks [43]. It should be noted that more scales can be added to the proposed network if the computation time is not of a concern. The Generator architecture has the following sequence of blocks:

Half Scale Part:
ConvBlock1(1,64)
ConvBlock2(128)-ConvBlock2(256)-
ConvBlock2(512)-ConvBlock2(1024)- SelfAttentionBlock,
ResBlock(1024)× 9,
ConvBlock3(512)-ConvBlock3(256)-
ConvBlock3(128)-ConvBlock3(64)-
ConvBlock1(1,1)-SelfAttentionBlock.

Full-Scale Part:
ConvBlock1(1,32)-ConvBlock2(64)-
(Output of this is added with the self attention maps from the Half Scale part.)
ResBlock(64)× 3,
ConvBlock3(32)-ConvBlock1(1,1).

The discriminator architecture has the following sequence of blocks:
ConvBlock(64),ConvBlock(128),
ConvBlock(256),ConvBlock(512).

The details about the layers in each of the above blocks is explained in the appendix. The overall network architecture is illustrated in Fig 4. We perform certain geometric transformations like translation, rotation and resizing of the segmentation masks to obtain new segmentation maps that have the possibility of existing in the real world. The intensity of these transformations are decided after making careful evaluations on the resultant maps such that they do not represent anything that is impossible to occur in real time. This helps us produce a numerous amount of new segmentation masks and their corresponding ultrasound images (using our MSSA Network), thus introducing new data into the training set which enhances the generalization of segmentation network.

B. Loss for MSSA Network

Let G denote the generator network and D_1, D_2 denote the discriminator networks. Our objective function to train the overall network is as follows

$$\min_G((\max_{D_1,D_2} \sum_{k=1,2} \mathcal{L}_{GAN}(G, D_k)) + \lambda_1 \sum_{k=1,2} \mathcal{L}_{FM}(G, \mathcal{D}_k)), \tag{6}$$

where

$$\begin{aligned} \mathcal{L}_{\mathcal{GAN}}(G, D_k) = \mathbb{E}_{(x,s)} \log D_k(x, s) \\ + \mathbb{E}_x[(\log(1 - D_k(x, G(x)))], \end{aligned} \tag{7}$$

and

$$\mathcal{L}_{\mathcal{FM}}(G, D_k) = \mathbb{E}_{(x,s)} \sum_{i=1}^{T} \frac{1}{N_i} \|[\log D_k^i(x, s) - D_k^i(x, G(x))]\|_2^2$$

are the two loss functions in the overall objective function. It can be noted that x is the US scan that is to be synthesized and s is the input segmentation mask. $\mathcal{L}_{GAN}$ is the standard GAN loss which is the sum of expectation over the discriminator's estimate of how much probability that the data instance is real/fake depending on whether it is a real data instance or if it is synthesized from the generator. $\mathcal{L}_{FM}$ is the feature matching loss which is a perceptual loss [44] calculated across different layers in the discriminator. λ_1 is the factor which controls the amount of feature matching loss that is to affect objective function. N_i denotes the number of elements in the i^{th} layer and T denotes the total number of layers in the network.

IV. Experiments and Results

In this section, we present details of the experiments and quality measures used to evaluate the proposed synthesis and segmentation methods. We also discuss the dataset and training details followed by comparison of the proposed methods against a set of baseline methods and recent state-of-the-art approaches.

A. Dataset

After obtaining institutional review board (IRB) approval, retrospective brain US scans were collected. A total of 1629 in vivo B-mode US images were obtained from 20 different subjects (age<1 years old) who were treated between 2010 and 2016. The dataset contained subjects with IVH and without (healthy subjects but in risk of developing IVH). The US scans were collected using a Philips US machine with a C8-5 broadband curved array transducer using coronal and sagittal scan planes. For every collected image ventricles and septum pellecudi are manually segmented by an expert ultrasonographer. We split these images randomly into 1300 Training images and 329 Testing images for experiments. Note that these images are of size 512×512. During the random split of the dataset the training and testing data did not include the same patient scans. Sample images and the corresponding segmentation masks from this dataset are shown in Fig. 6. As there are more variations in each ultrasound used, our dataset does not have much potential for over-fitting.

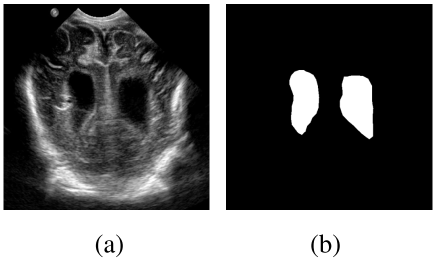
(a) (b)

Fig. 6. Sample brain ultrasound image from the dataset. (a) Brain US image. (b) the corresponding segmentation mask.

We evaluate the performance of both our segmentation and synthesis methods with recent methods on the randomly selected 329 test images. We compare the performance of our synthesis method against, pix2pix [5], U-Net [6], UDe-Net [40], and Wang et al. [7]. We conduct these experiments three times and average out the obtained results. We use DICE coefficient and Intersection over Union (IoU) to measure the performance of different segmentation networks. For the synthesis network, we compare our method with pix2pix [5], Self-Attention GAN [41] and pix2pixHD [8]. We calculate the DICE accuracy of CBAS when trained with the synthetic data generated from each of the compared methods to validate the performance of our method apart from the qualitative results.

B. Training Details

CBAS is trained using $\mathcal{L}_{final}$ with the Adam optimizer [45] and batch size of 1. The learning rate is set equal to 0.0002 and annealed by 5% for every 10 epochs. λ is set equal to 0.1 for initial epochs, but when the mean of all values in the confidence maps $c_{\times 1}, c_{\times 2}$ is greater than 0.75 then λ is set equal to 0.01. CBAS is trained for 100 epochs. We perform data augmentation using horizontal, vertical flips and random crops to extend the training images to 6500 images. We resize the images to 640×640 and crop 512×512 size patches to obtain random crop images.

MSSA is trained using a learning rate of 0.0002 with the Adam optimizer [45] and batch size of 1. The half-scale self-attention guided part of the generator is trained separately for the first 200 epochs. Then, the full scale part of the network is trained along with this for the next 300 epochs. λ_1 in Eq. (6) is set equatl to 0.1.

C. Qualitative Performance

Fig. 5 shows the qualitative performance of different segmentation methods on the test images. We can clearly see that pix2pix [5], U-Net [6], UDe-Net [6], [40], and Wang et al. [7], misclassified normal regions as the brain ventricular regions. For example, from the second column of Fig. 5, we can clearly observe under segmentation of brain ventricles regions in the outputs produced using pix2pix [5]. Brain ventricle segmentations obtained using U-Net [6], and UDe-Net [40] also contain under segmentation for large size ventricles (in the fourth row) and over segmentation for small size ventricles as shown in the third and the fourth columns of Fig. 5. Wang et al. [7] produce brain segmentations which contain inaccurate edges for large ventricles and under segmentation for small size ventricles. On the other hand, the estimated shape of the brain ventricular regions by those methods are slightly off when compared to the original shape. Visually we can see that CBAS produces more accurate brain ventricular regions, and does not miss-classify the normal regions as brain ventricular regions.

Fig. 8 shows the qualitative performance of different synthesis methods. We observe that pix2pix [5] is very unstable at generating high-resolution images and performs very poorly in almost every case. The pix2pixHD method [8] synthesizes high-resolution images but fails to synthesize realistic looking images. In the second row, it can be observed that pix2pixHD does not properly capture the features of the US image near the edges. Similarly, as can be seen from the third, fourth and fifth rows, the structures inside the US image are not captured by pix2pixHD. Our proposed method captures all these structures that are missed by pix2pixHD which can be seen in the illustration.

D. Quantitative Performance

Table. I shows the quantitative performance of our proposed segmentation method and the other investigated methods. As it can be seen from this table, our method clearly outperforms these recent segmentation methods (p<0.05 for paired t-test with 5% significance). The paired t-test value using the DICE scores between CBAS and Wang et.al[7] (second best method), resulted in an average p value of 3.63×10^{-9}. Note that, our method has very less number of parameters in the network as compared to Wang et al. [7]. Time taken by our method to process an image of 512×512 is about 0.01 seconds compared to 0.02 seconds for [7]. This presents a 50% improvement in computation time.

Table III shows the quantitative performance of our proposed synthesis method and the other recent methods. The DICE accuracy is calculated by training our proposed segmentation method (CBAS) on equal proportion of real and synthetic images, where the synthetic images are generated by the methods we compare. We use a total of 2600 (1300 real and 1300 synthetic) images to train our CBAS network. Also, the number of synthetic data that can be synthesized once MSSA is trained is indefinite as minor transformations in the segmentation mask like translation, resizing of the mask can be done to obtain a new ground truth. In our work we have

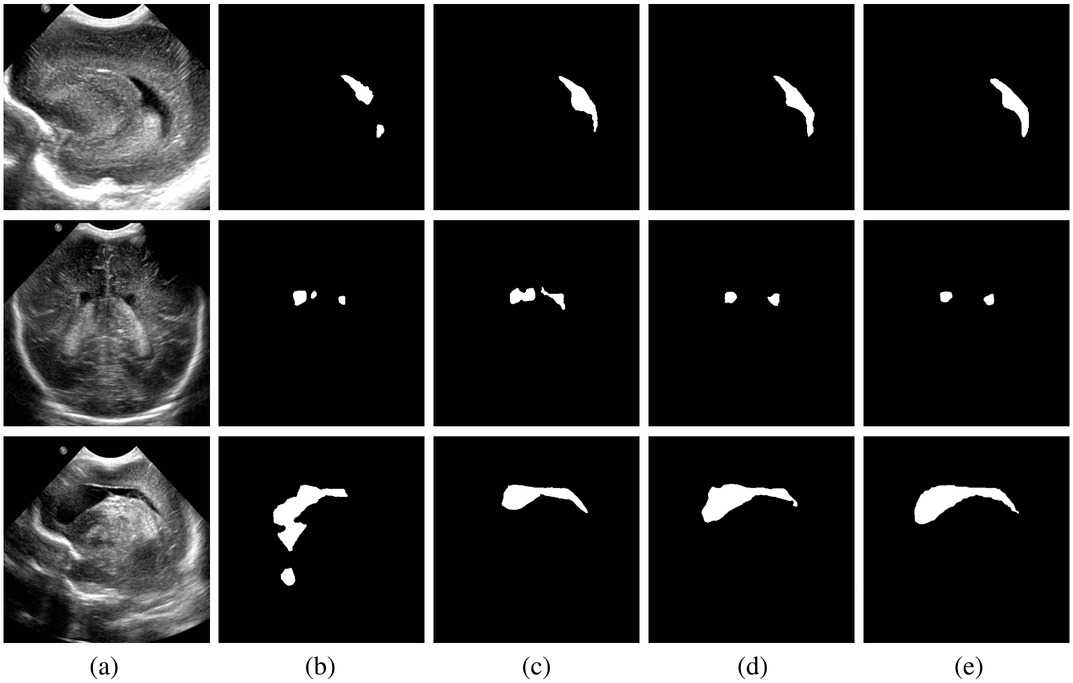

Fig. 7. Qualitative results on test images for ablation study. (a) Input brain ultrasound image, (b) BN, (c) BN w/ SB, (d) CBAS (ours), (e) Ground-truth ventricle segmentation. Brain ventricular segmentations from BN and BN w/ SB have over segmentation for small size ventricles, and incorrect segmentation at the edges of large size ventricles. CBAS trained with $\mathcal{L}_{final}$ produced best results with accurate segmentation for both small and large sized ventricles.

TABLE II
QUANTITATIVE RESULTS CORRESPONDING TO ABLATION STUDY. RESULTS SHOWN CORRESPOND TO MEAN VALUES.

Method	Loss	DICE	IoU(%)	p-value
BN	CE	0.8538	77.90	6.48 $\times 10^{-3}$
BN w/ SB	CE	0.8673	78.48	3.34 $\times 10^{-2}$
BN w/ SB and CB	CE	0.8664	78.56	4.74 $\times 10^{-2}$
CBAS	$\mathcal{L}_{final}$	0.8813	80.25	–
CBAS (with synthetic data)	$\mathcal{L}_{final}$	**0.8901**	**81.03**	–

fixed the number of synthetic data generated to be always equal to the number of real data. We train the mixture of real and synthetic data for half the number of epochs we train with only real data for fair comparison. From Table III, it can be observed that the segmentation network performs the best when the synthesized images added are generated using our proposed method. It can be noted that the addition of self attention to the base network [8] improved the qualitative and quantitative results as seen in Table III and Fig 8.

TABLE III
COMPARISON OF DIFFERENT IMAGE SYNTHESIS METHODS IN TERMS OF DICE (SEGMENTATION PERFORMANCE OF CBAS WHEN TRAINED ON A MIXTURE OF REAL AND SYNTHETIC IMAGES, SYNTHESIZED USING METHODS WHICH ARE COMPARED)

Method	DICE Accuracy (%)
pix2pix[5]	80.12
SA-GAN[41]	83.41
pix2pixHD[8]	86.23
MSSA (ours)	**89.01**

We conduct further experiments to ascertain the importance of the synthetic data that is generated. Table IV contains DICE accuracies of the CBAS network when trained with different proportions of the real data. The total number of images on which the images are trained are always 1300 in every case. The percentage of real data out of the 1300 is different for every case except for the 100% case. For example when the network is trained with 50% real data, it is trained with 650 images. When it is trained with 50% real and synthetic data, it is trained with 650 real images and 650 synthetic images. Only in the 100% case, the number of images used for the real case is 1300 and the number of images used for real with synthetic data is 2600 images. Real+Syn (ST-Ratio) column corresponds to the results where the Synthesis network was also trained with ratios of data same as that of segmentation network was trained on. ST-Full corresponds to the results where synthesis network was trained with full data. It can be seen from the table that the addition of synthetic data is highly useful in cases where the real data availability is very low. Also, even when the network is trained only on the synthetic data, it gives a dice accuracy of 83.42%. We also illustrate the performance gap produced by adding synthetic

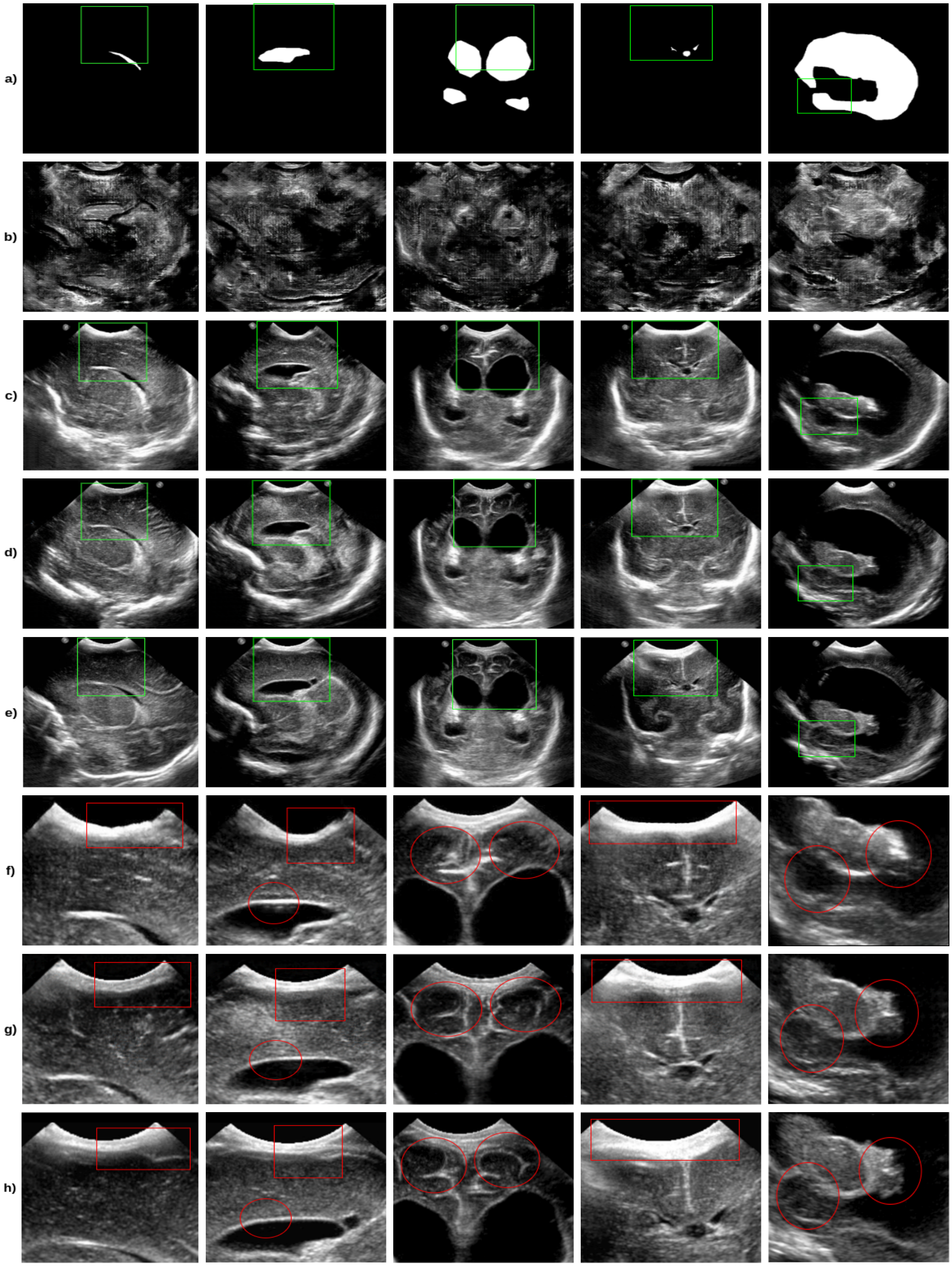

Fig. 8. Qualitative results on test images for the synthesis task. (a) Input segmentation mask, (b) Synthesized image using pix2pix [5], (c) Synthesized image using pix2pixHD [8], (d) Synthesized image using MSSA (ours), (e) Real B-Mode Ultrasound image for the input segmentation mask in (a). Images shown in (f),(g) and (h) are the zoomed in parts inside the green box as shown in (c),(d) and (e) respectively. The red boxes in (f),(g) and (h) denote the specific structures that show how our method is closer to the real image than pix2pixHD [8].

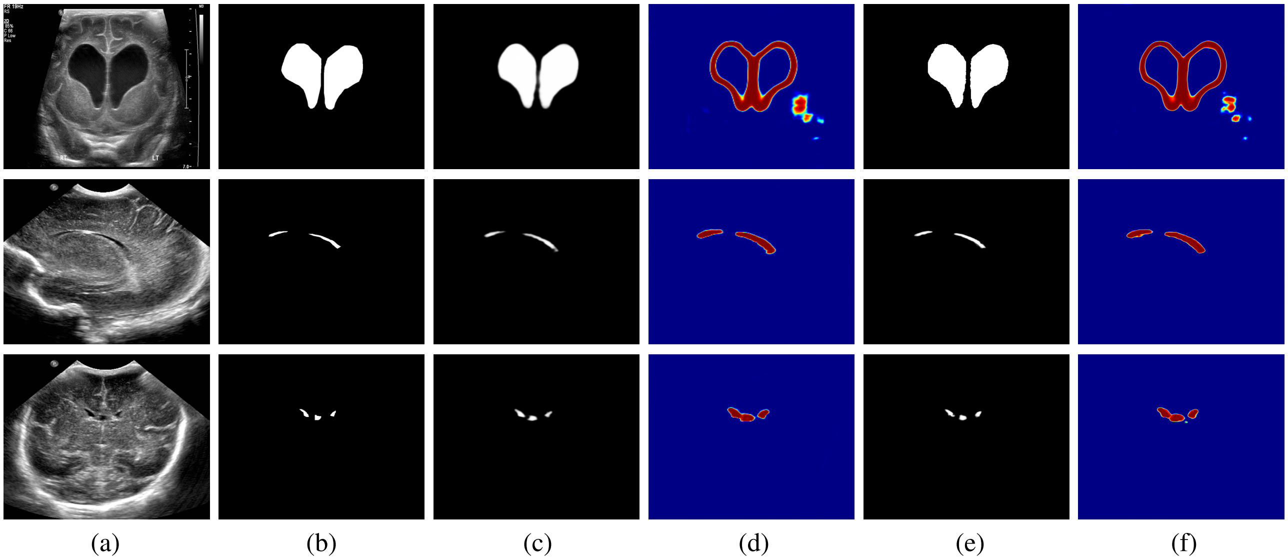

Fig. 9. Confidence maps visualization for a test image. (a) Input brain ultrasound image. (b) Ground-truth ventricle segmentation. (c), (e) are the estimated vetricle segmentations at different scales $\hat{s}_{\times 2}$, and $\hat{s}_{\times 1}$, respectively. (d),(f) are the corresponding confidence maps $c_{\times 2}$, and $c_{\times 1}$, respectively. Note that in the confidence maps blue means 1 and red means 0

data in Fig 10 where we can see how synthetic data helps when the real data is very less. It should also be noted that in all these experiments (except in 100% case), the upper bound of the number of training data was fixed as 1300 to give a reasonable comparison. An interesting point to be noted in this comparison is that our proposed method always outperforms the normal baseline even if it is trained with less number of data than the normal baseline. For example, from Table IV, it can be seen that when we train CBAS with 50 % data, the dice accuracy is 83.61. When we train CBAS with mixture of 25 % real and synthetic data, we get 87.43 % for ST-Full and 84.63 % for ST-Ratio, both of which are larger than 50 % data protocol. This trend can be observed across all ratios in the Table IV. So our proposed method actually works better than the baseline even if it is trained with half of the real data as used to train the baseline.

E. Ablation Study

We study the performance of each block's contribution to CBAS by conducting various experiments on the test images. We start with the UNet base network (BN), and then add SB blocks to estimate the segmentation maps at different scales. Finally, we add CB block to construct CBAS and train it with $\mathcal{L}_{final}$. Table. II shows the contribution of each block on the CBAS network. Note that BN and BN w/ SB are trained using the cross-entropy (CE) loss. The base network, BN itself produces poor results. However, when SB blocks are added to BN, the performance improves significantly. The combination of BN, SB and CB to construct CBAS and trained with $\mathcal{L}_{final}$ produces the best results. Table. II clearly shows the importance of formulating ML inference and training CBAS with $\mathcal{L}_{final}$. This can be clearly seen by comparing the performance of CBAS when trained with and without $\mathcal{L}_{final}$. We computed the p values using the DICE scores for the results obtained after adding different components to base network to obtain CBAS, against the DICE scores for the final results obtained using CBAS shown in the Table. II.

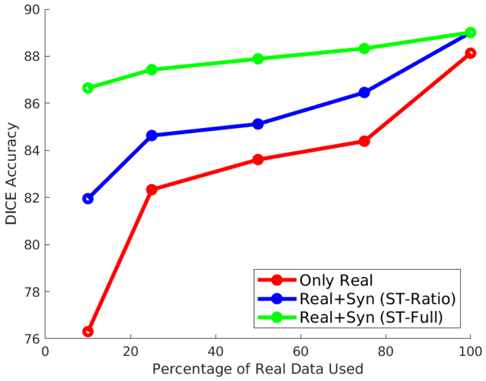

Fig. 10. Performance of the segmentation network while trained on different proportions of the real data. Real+Syn (ST-Ratio) corresponds to performance of segmentation network trained on a mixture of real and synthetic data where the synthetic data is trained with ratios of the data, and the segmentation network is also trained with ratios of data. Real+Syn (ST-Full) corresponds to performance of segmentation network trained on a mixture of real and synthetic data where the synthetic data is trained on full data, and the segmentation network is trained with ratios of data.

Fig. 7 shows the qualitative performance of BN, BN w/SB, and CBAS. We can clearly see the progressive improvements visually when each block is added to BN. For example in the first column of Fig. 7, the output brain ventricular segmentation regions are random at the edges for large size ventricles, and contains over segmentation of normal regions for small size ventricles. Once we add the SB blocks to BN, the outputs get much better compared to BN, but we can still observe some under segmentations in large ventricles and over segmentation in small size ventricles as shown in the third column of Fig 7. Finally, when we add the CB blocks

TABLE IV
COMPARISON OF THE SEGMENTATION PERFORMANCE OF CBAS ACROSS DIFFERENT PROPORTIONS OF THE REAL DATA.

% Real	Only Real	Real+Syn (ST-Ratio)	Real+Syn (ST-Full)
100 %	0.8813	0.8901	0.8901
75 %	0.8439	0.8646	0.8833
50 %	0.8361	0.8512	0.8789
25 %	0.8233	0.8463	0.8743
10 %	0.7630	0.8195	0.8665
0 % (only synthetic)	-	-	0.8342

to construct CBAS and train it with $\mathcal{L}_{final}$, we observe the best results as shown in the fourth column of Fig. 7. Final outputs have clear edges for larger ventricles and accurate segmentation for smaller size ventricles.

Fig. 9 shows brain ventricle segmentation, and the corresponding confidence maps at different scales. We clearly observe $c_{\times 2}$, $c_{\times 1}$ (fourth and sixth columns in Fig. 9 respectively) highlight the erroneous regions $\hat{s}_{\times 2}$, $\hat{s}_{\times 1}$ (third and fifth columns in Fig. 9 respectively) which guide the CBAS to learn the accurate segmentation in those regions. For example, as shown in Fig. 9, the edges of the brain ventricle segmentation are highlighted in the confidence maps by producing low confidence scores using the CB blocks. This makes CBAS more attentive in those regions while calculating the segmentation maps.

V. CONCLUSION

We proposed a novel method, called CBAS, to address the US brain anatomy segmentation task. In our approach, we introduced a technique to estimate segmentation and the corresponding confidence maps. Additionally, we trained our CBAS network with proposed novel loss function $\mathcal{L}_{final}$. Extensive experiments showed that CBAS outperformed the state-of-the-art methods with fewer number of parameters. The reported computational time makes CBAS the best match for real-time applications. On top of that, we proposed a image synthesis method to add synthetic data to our training data, which further boosts the performance of CBAS. We also show from various experiments that our proposed synthesis method is better than recent methods.

Although, our proposed method outperforms state-of-the-art methods, several limitations in our study still exists. First our method is geared towards segmenting 2D US data. 2D scans are inherently limited to cross-sectional analysis and do not take advantage of surface continuity between adjacent images (i.e., along the axis perpendicular to the scan plane direction). Currently, we are in the process of collecting 3D US scans. Therefore, in the future, we will extend our method for processing volumetric US data. Second limitation is related to the the fact that manual segmentation, performed by single expert ultrasonographer with more than 20 years of experience, was treated as gold standard in our study. Due to the typical US imaging artifacts manual segmentation of US data is an error prone process. Shape of the anatomical region to be segmented and expertise of the ultrasonographer will bias the obtained segmentation results. Future work will also involve the investigation of inter- and intra-user variability of the segmentation and its effect on the proposed method. During monitoring of the preterm neonates in situations where the diagnosis can not be assessed with US additional imaging using MRI is performed. Anatomical structures segmented from MRI data could be treated as a gold standard segmentation to minimize the variability of manual segmentation from US data. Unfortunately, none of the enrolled subjects had an MRI scan available. Therefore this analysis could not be performed during this work. We also did not calculate any quantitative US measurements such as ventricular index (VI), anterior horn width (AHW), and thalamo-occipital distance (TOD). These measurements are usually calculated manually from B-mode US data [46]. In or future work we will extend out network for simultaneous segmentation and anatomical landmark extraction in order to automate the quantitative measurement process. Finally, during this work we have only focused on lateral ventricles and septum pellecudi. Segmentation of third and fourth ventricles were beyond the scope of this study. However, quantitative measures obtained from these ventricles should be considered as a valuable additional information to evaluate the pathophysiology of ventriculomegaly [47].

APPENDIX A
DETAILS OF DIFFERENT BLOCKS IN CBAS

Table V shows the details regarding ResBlock, Segmentation Block and Confidence Block in our network. Note that in Table V C, H and W denote the number of channels, height and width of the intermediate feature maps respectively.

TABLE V
CONFIGURATION OF BLOCKS IN THE CBAS NETWORK.

Block name	Layer	Kernel size	Filters	dilation	Input size	Output size
ResBlock	Conv1	1 x 1	2C	1	C × H × W	2C × H × W
	Conv2	3 × 3	2C	1	2C × H × W	2C × H × W
	Conv3	3 × 3	C	2	2C × H × W	C × H × W
Segmentation Block	Conv1	1 × 1	32	1	64 × H × W	32 × H × W
	Conv2	3 × 3	32	1	32 × H × W	32 × H × W
	Conv3	3 × 3	16	1	32 × H × W	16 × H × W
	Conv4	3 × 3	1	1	16 × H × W	1 × H × W
Confidence Block	Conv1	1 × 1	16	1	33 × H × W	16 × H × W
	Conv2	3 × 3	16	1	16 × H × W	16 × H × W
	Conv3	3 × 3	16	1	16 × H × W	16 × H × W
	Conv4	3 × 3	1	1	16 × H × W	1 × H × W
	Sigmoid	–	–	–	1 × H × W	1 × H × W

APPENDIX B
DETAILS OF MSSA NETWORK

A. Generator

Table VI shows the details of each block in the generator network's architecture. Note that, k is the number of filters in the convolutional layers in blocks, where ever specified. C is the number of channels of input fed into the convolutional layer in the blocks, where ever specified.

TABLE VI
CONFIGURATION OF THE SYNTHESIS GENERATOR NETWORK.

Block name	Layer	Kernel size	Filters	Stride	Input size	Output size
ConvBlock 1	Conv1	7 × 7	k	1	1 × H × W	k × H × W
	InstanceNorm	–	–	–	k × H × W	k × H × W
	ReLU	–	–	–	k × H × W	k × H × W
ConvBlock 2	Conv1	3 × 3	k	2	C × H × W	k × H/2 × W/2
	InstanceNorm	–	–	–	k × H/2 × W/2	k × H/2 × W/2
	ReLU	–	–	–	k × H/2 × W/2	k × H/2 × W/2
ConvBlock 3	Conv1	3 x 3	k	0.5	C × H × W	k × 2H × 2W
	InstanceNorm	–	–	–	k × 2H × 2W	k × 2H × 2W
	ReLU	–	–	–	k × 2H × 2W	k × 2H × 2W
Self Attention Block	Query-Conv1	1 x 1	128	1	C × H × W	128 × H × W
	Key-Conv2	1 × 1	128	1	128 × H × W	128 × H × W
	Value-Conv3	1 × 1	1024		128 × H × W	1024 × H × W
ResBlock	Conv1	3 × 3	k	1	C × H × W	k × H × W
	Conv2	3 × 3	k	1	k × H × W	k × H × W

B. Discriminator

Table VII shows the details of each block in the discriminator's network architecture. Note that, k is the number of filters in the convolutional layers in the block.

TABLE VII
CONFIGURATION OF BLOCKS IN THE DISCRIMINATOR NETWORK.

Block name	Layer	Kernel size	Filters	Stride	Input size	Output size
ConvBlock	Conv1	4 x 4	k	1	1 × H × W	k × H × W
	InstanceNorm	–	–	–	k × H × W	k × H × W
	LeakyReLU	–	–	–	k × H × W	k × H × W

REFERENCES

[1] H. Blencowe, S. Cousens, D. Chou, M. Oestergaard, L. Say, A.-B. Moller, M. Kinney, and J. Lawn, "Born too soon: the global epidemiology of 15 million preterm births," *Reproductive health*, vol. 10, no. 1, p. S2, 2013.

[2] S. Robinson, "Neonatal posthemorrhagic hydrocephalus from prematurity: pathophysiology and current treatment concepts: a review," *Journal of Neurosurgery: Pediatrics*, vol. 9, no. 3, pp. 242–258, 2012.

[3] D. M. Sherer, M. Sokolovski, M. Dalloul, P. Santoso, J. Curcio, and O. Abulafia, "Prenatal diagnosis of dilated cavum septum pellucidum et vergae," *American journal of perinatology*, vol. 21, no. 05, pp. 247–251, 2004.

[4] M. Sarwar, "The septum pellucidum: normal and abnormal." *American Journal of Neuroradiology*, vol. 10, no. 5, pp. 989–1005, 1989.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[7] P. Wang, N. G. Cuccolo, R. Tyagi, I. Hacihaliloglu, and V. M. Patel, "Automatic real-time cnn-based neonatal brain ventricles segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 716–719.

[8] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[9] W. Qiu, Y. Chen, J. Kishimoto, S. de Ribaupierre, B. Chiu, A. Fenster, and J. Yuan, "Automatic segmentation approach to extracting neonatal cerebral ventricles from 3d ultrasound images," *Medical image analysis*, vol. 35, pp. 181–191, 2017.

[10] M.-A. Boucher, S. Lippé, A. Damphousse, R. El-Jalbout, and S. Kadoury, "Dilatation of lateral ventricles with brain volumes in infants with 3d transfontanelle us," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 557–565.

[11] B. Sciolla, M. Martin, P. Delachartre, and P. Quetin, "Segmentation of the lateral ventricles in 3d ultrasound images of the brain in neonates," in *2016 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2016, pp. 1–4.

[12] M. Martin, B. Sciolla, M. Sdika, X. Wang, P. Quetin, and P. Delachartre, "Automatic segmentation of the cerebral ventricle in neonates using deep learning with 3d reconstructed freehand ultrasound imaging," in *2018 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2018, pp. 1–4.

[13] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1442–1450.

[14] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

[15] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] R. Mehta, T. Christinck, T. Nair, P. Lemaitre, D. Arnold, and T. Arbel, "Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*. Springer, 2019, pp. 23–32.

[17] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020.

[18] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 48–56.

[19] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 605–613.

[20] J. Merkow, A. Marsden, D. Kriegman, and Z. Tu, "Dense volume-to-volume vascular boundary detection," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 371–379.

[21] S. Gur, L. Wolf, L. Golgher, and P. Blinder, "Unsupervised microvascular image segmentation using an active contours mimicking neural network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 722–10 731.

[22] A. Hatamizadeh, D. Terzopoulos, and A. Myronenko, "End-to-end boundary aware networks for medical image segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 187–194.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[26] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 417–425.

[27] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Fulham, "Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans)," in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*. Springer, 2017, pp. 43–51.

[28] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "Gan-based synthetic brain mr image generation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 734–738.

[29] Q. Yang, N. Li, Z. Zhao, X. Fan, E.-C. Chang, Y. Xu *et al.*, "Mri image-to-image translation for cross-modality image registration and segmentation," *arXiv preprint arXiv:1801.06940*, 2018.

[30] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, p. 101684, 2019.

[31] H. Zhao, H. Li, S. Maurer-Stroh, and L. Cheng, "Synthesizing retinal and neuronal images with generative adversarial nets," *Medical image analysis*, vol. 49, pp. 14–26, 2018.

[32] O. Bailo, D. Ham, and Y. Min Shin, "Red blood cell image generation for data augmentation using conditional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[33] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Capsulegan: Generative adversarial capsule network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[34] T. Fujioka, M. Mori, K. Kubota, Y. Kikuchi, L. Katsuta, M. Adachi, G. Oda, T. Nakagawa, Y. Kitazume, and U. Tateishi, "Breast ultrasound image synthesis using deep convolutional generative adversarial networks," *Diagnostics*, vol. 9, no. 4, p. 176, 2019.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[36] Y. Hu, E. Gibson, L.-L. Lee, W. Xie, D. C. Barratt, T. Vercauteren, and J. A. Noble, "Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks," in *Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment*. Springer, 2017, pp. 105–115.

[37] F. Tom and D. Sheet, "Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1174–1177.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization."

[40] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[41] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[44] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] M. Davies, M. Swaminathan, S. Chuang, and F. Betheras, "Reference ranges for the linear dimensions of the intracranial ventricles in preterm neonates," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 82, no. 3, pp. F218–F223, 2000.

[47] M. J. Brouwer, L. S. de Vries, F. Groenendaal, C. Koopman, L. R. Pistorius, E. J. Mulder, and M. J. Benders, "New reference values for the neonatal cerebral ventricles," *Radiology*, vol. 262, no. 1, pp. 224–233, 2012.

Jeya Maria Jose is Ph.D. student in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University. Prior to joining Hopkins, he graduated from NIT Trichy with a Bachelors degree in Instrumentation and Control Engineering. He also spent some time working in the Biomedical Engineering Department at National University of Singapore (NUS) as a visiting research intern. His research interests include deep learning based solutions to medical image segmentation, image synthesis, image restoration and computer vision.

Rajeev Yasarla is PhD student in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University. Prior to joining Hopkins, he graduated from IIT Madras with Bachelors and Masters degree in Electrical Engineering. His research interests include deep learning based image restoration (like de-raining, deblurring, and atomspheric turbulence distortion removal), object detection and medical image segmentation.

Puyang Wang is PhD student in the Department of Electrical and Computer Engineering at Johns Hopkins University. Prior to joining Hopkins, he graduated from University of Electronic Science and Technology of China with Bachelors degree in Electrical Engineering in 2016. His research interests include deep learning based image restoration and medical image analysis including classification, segmenation and reconstruction.

Ilker Hacihaliloglu received his B.Sc and M.Sc. degrees in Electrical Engineering from Istanbul Technical University (Istanbul, Turkey) in 2001 and 2004 respectively. He completed his PhD degree in Electrical and Computer Engineering from the University of British Columbia (Vancouver, Canada) in 2010. He has performed his Post-doctoral studies at the Vancouver General Hospital and Center for Hip Health and Mobility. He is a Assistant Professor in the Department of Biomedical Engineering, Rutgers University. His specific research focus is on the development of ultrasound-based surgical and diagnostic systems.

Vishal M. Patel [SM'15] is an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University. Prior to joining Hopkins, he was an A. Walter Tyson Assistant Professor in the Department of ECE at Rutgers University and a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). He completed his Ph.D. in Electrical Engineering from the University of Maryland, College Park, MD, in 2010. His current research interests include signal processing, computer vision,and pattern recognition with applications in biometrics and imaging. He has received a number of awards including the 2016 ONR Young Investigator Award, the 2016 Jimmy Lin Award for Invention, A. Walter Tyson Assistant Professorship Award, Best Paper Award at IEEE AVSS 2017 & 2019, Best Paper Award at IEEE BTAS 2015, Honorable Mention Paper Award at IAPR ICB 2018, two Best Student Paper Awards at IAPR ICPR 2018, and Best Poster Awards at BTAS 2015 and 2016. He is an Associate Editor of the IEEE Signal Processing Magazine, Pattern Recognition Journal, and serves on the Machine Learning for Signal Processing (MLSP) Committee of the IEEE Signal Processing Society. He serves as the vice president of conferences for the IEEE Biometrics Council. He is a member of Eta Kappa Nu, Pi Mu Epsilon, and Phi Beta Kappa.