D2ANET: DENSELY ATTENTIONAL-AWARE NETWORK FOR FIRST TRIMESTER ULTRASOUND CRL AND NT SEGMENTATION

Mourad Gridach[1] *Robail Yasrab*[1] *Lior Drukker* [2] *Aris T. Papageorghiou*[2] *J. Alison Noble*[1]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[2]Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK

ABSTRACT

Manual annotation of medical images is time consuming for clinical experts; therefore, reliable automatic segmentation would be the ideal way to handle large medical datasets. In this paper, we are interested in detection and segmentation of two fundamental measurements in the first trimester ultrasound (US) scan: Nuchal Translucency (NT) and Crown Rump Length (CRL). There can be a significant variation in the shape, location or size of the anatomical structures in the fetal US scans. We propose a new approach, namely Densely Attentional-Aware Network for First Trimester Ultrasound CRL and NT Segmentation (DA2Net), to encode variation in feature size by relying on the powerful attention mechanism and densely connected networks. Our results show that the proposed D2ANet offers high pixel agreement (mean JSC = 84.21) with expert manual annotations.

Index Terms— Video Segmentation, First Trimester, Spatial Attention, Channel Attention, Ultrasound.

1. INTRODUCTION

The goal of the scan is to ensure pregnancy viability, establish accurate pregnancy dating, and assess the risk for chromosomal anomalies [1]. A sonographer carries out the first-trimester scan and manipulates the US probe to achieve a series of standard imaging planes, which is subjected to comprehensive training and experience [2]. Therefore, an automatic fetal anatomy detection and overlaying method could offer assistance and support to trainees and newly qualified operators. It has been demonstrated that semantic segmentation-based methods [3, 4] can accurately recognize and analyze complex fetal anatomy in US video scans.

Most of the existing semantic segmentation approaches rely on the encoder-decoder structure [4, 5, 6], which suffer from a fundamental issue, namely the loss of spatial interconnection due to the use of consecutive pooling layers and strided convolutions. Other approaches use spatial attention, channel attention or a combination of both either in a parallel or sequential manner. Although they have achieved improved performance, they suffer from two main issues: (1) directly fusing the spatial and channel attention feature maps may produce incorrect importance weights for pixel representations, and (2) it is unclear what will be the contributions of each attention.

Related Work. There are a few works that have studied first-trimester US scans such as standard plane detection [7], segmentation [8], and fetal biometry [9]. Most of these methods [10, 8] employed traditional encoder-decoder designs for semantic segmentation. The majority of these methods use traditional encoder-decoder designs that require manual annotation of medical images, creating fundamental bottlenecks in the process. Recently, attention-based methods relying on spatial attention and channel attention or combining both are gaining more popularity to resolve aforementioned bottleneck. Hu et al., [11] developed an architecture based on channel attention to examine the relationships between channels to increase the most important among them, which improves the model's performance. DANet [12] proposed a deep learning architecture that combines spatial and channel attention by directly adding them together to model the dependencies along both dimensions (position and channel). Although these approaches capture dependencies in all dimensions, they still treat the dimensions separately, which can cause conflicting results. To the best of our knowledge, we make the first attempt at fusing channel and spatial attentions in better way and combine them with dense connectivity for first trimester US fetal NT and CRL segmentation.

Contribution. This paper proposes a novel network called Densely Attentional-Aware (D2ANET) Network that is trained and tested on First Trimester US video scans. The proposed design of D2ANET is capable to handle variations in feature size, location, and shape, which are the main challenges in fetal US images. D2ANet low-compute design is 18% faster than CE-Net [6]. It uses (1) channel-fusion and spatial attention-based novel Pair Attention Block (PAB), and (2) dense connectivity to captures important contextual features, which improves the overall segmentation performance. The experimental results show the effectivness of the proposed D2ANet model on two (NT and CRL) first-trimester US segmentation datasets.

Clinical data acquisition was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee.

Full-length First Trimester Ultrasound Scans			
CRL Task 128 Subjects/Videos		NT Task 146 Subjects/Videos	
Training	Testing	Training	Testing
Videos=100 (Frames=12,534) 77.9%	Videos=28 (Frames=3,559) 22.1%	Videos=110 (Frames=10,174) 79.3%	Videos=36 (Frames=2,647) 20.7%

Fig. 1: Details of datasets and tasks used in this study

2. OUR METHOD

We propose a novel Densely Attentional-Aware Network trained and tested on first-trimester ultrasound (US) scan video dataset for automatic segmentation CRL and NT. It contains two major ideas: a novel Pair Attention Block (PAB) with a Spatio-Channel Fusion layer and a dense connectivity. On the one hand, the PAB block with the SCF layer enable the model to fuse the channel and spatial attentions in proper way to handle both the variations in NT and CRL structures sizes in fetal US images and capture the appropriate contextual information. On the other hand, inspired by the success of dense connectivity networks, we add dense connections to capture the advantages of DenseNets.

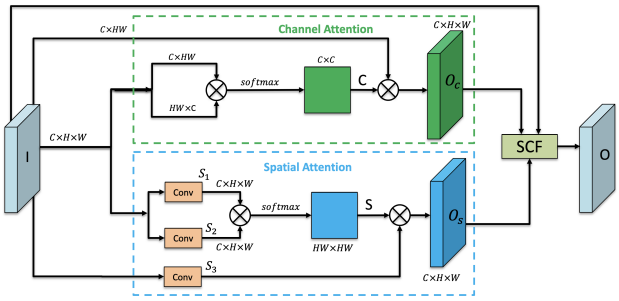


Fig. 2: Pair Attention Block architecture consists of fusing channel attention and spatial attention

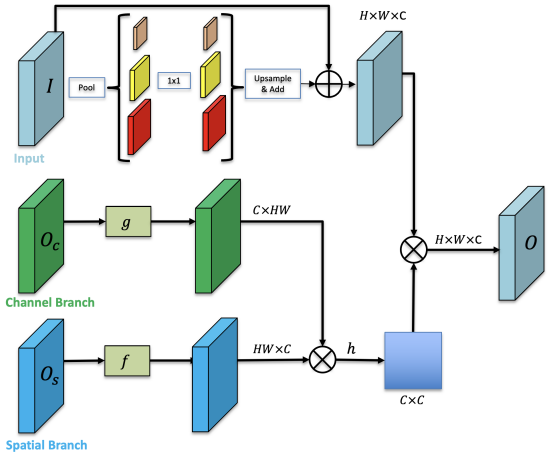


Fig. 3: Spatio-Channel Fusion architecture

2.1. Pair Attention Block and Spatio-Channel Fusion

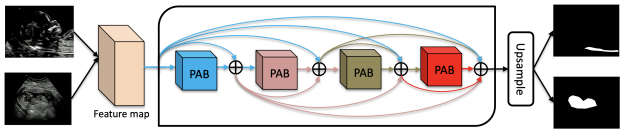


Fig. 4: D2ANet architecture consists of dense connectivity and PAB blocks

In order to address the issues of feature combination in the existing channel and spatial attention approaches, we designed a novel Pair Attention Block (PAB), illustrated in Fig. 2, which better combines the advantages of channel attention and spatial attention. Given an input feature map, we apply a channel attention and spatial attention in parallel, where the result is a feature map with the same dimension as the input. Given the input feature map $I \in \mathbf{R}^{C\times H\times W}$, the following equations summarize the spatial attention:

$$O_s^i = \sum_{j=1}^{N} S(i,j)\ S_3(j) + I_i$$
$$S(i,j) = softmax(S_1(i),\ S_2(j)) \tag{1}$$

where $N = H \times W$ and $S(i,j)$ measures the j^{th} position's impact on i^{th} position. The channel attention is summarized as follows:

$$O_c^i = \sum_{j=1}^{C} \big(C(i,j)\ I_j\big) + I_j$$
$$C(i,j) = softmax(I_i,\ I_j) \tag{2}$$

where $C(i,j)$ measures the j^{th} channel's impact on i^{th} channel. We feed the two feature maps to the Spatio-Channel Fusion (SCF) block, which will play the role of fusion. Fig. 3 shows the architecture of the SCF block (better shown in color). Segmenting the NT and CRL structures of the fetus could be challenging, therefore, it is essential for a network to detect such variations in location and feature size. Our SCF block solves this challenge by relying on its novel architecture design. Within SCF, we have three feature maps: the input, channel, and spatial feature maps. SCF is summarized as follows:

$$O = SCF(O_c, O_s, I) = h(O_c, O_s)\ pr(I) \tag{3}$$

where, $I \in \mathbf{R}^{C\times H\times W}$ is the output feature map and pr is the pooled representation applied to the input feature map I. First, for the spatial and channel attention feature maps, we apply a dot-product similarity:

$$h(O_c, O_s) = f(O_c)^T g(O_s) \tag{4}$$

where O_c and O_s are the the input feature maps from channel branch and spatial branch respectively. The two functions

$f(O_c) = W_f O_c$ and $g(O_s) = W_g O_s$ which are two embeddings, implemented using 1×1 convolution. We note that the function h plays an important role because it captures the feature similarity between the two attentions, which means that they contribute both for a better feature representation. Second, for the input feature map, it is divided into multiple pooled regions of different sizes (2×2, 3×3, 4×4 and 6×6), which are implemented and used in parallel. Lastly, a dot-product is used to get the final feature map.

By using the spatio-channel fusion block, our network captures different sized features. This is achieved using the dot-product between the pooled bins of various sizes and the two attentions (spatial and channel). Therefore, the network is able to cover the whole, half, or a small section of the image, and then fuse these to construct coherent information for the final NT and CRL segmentation.

2.2. D2ANet: Densely Attention-Aware Network

Relying on a single PAB block to extract useful features are not dense enough to cover the remarkable variations of NT and CRL structures (size, location, and shape) presented in fetal US images. Therefore, we propose to combine PAB block and dense connectivity to form the final D2ANet architecture. As shown in Fig. 4, each layer is connected to all subsequent layers in a feed-forward manner. This architecture has several advantages. First, it plays a role of regularization on datasets with small or medium sizes (such as medical image datasets), which reduces overfitting. Second, dense connectivity improves the flow of information and gradients over the network, which directly accelerates the training process. Third, feature reuse, which means that feature representation and gradient flow are improved after each layer. Lastly, an implicit deep supervision is applied through direct paths between all the feature maps.

3. EXPERIMENTS

3.1. Datasets and Baselines

Routine clinical first-trimester fetal US scans were available from a large-scale study PULSE. According to NHS Fetal Anomaly Screening Programme (FASP) guidelines [13] CRL and NT are two essential measurements in the first trimester US scan. US video was acquired through screen-grab signals at 30 frames per second of a GE Voluson E8 version BT18 (GE Healthcare, Zipf, Austria) US machine. On average a first-trimester US scan takes 13.73 ± 4.18 minutes, with an average of $24,720 \pm 7,534$ frames per scan video. The fetal structures CRL and NT mask were segmented using a training set of manually annotated video clips. Fig. 1 shows the details of each dataset. We compare our D2ANet with the following approaches: U-Net [5], CE-Net [6], and DANet [12].

Table 1: Comparison with state-of-the-art on NT dataset.

Methods	Acc	DSC	JSC	Sen
UNet	87.21 ± 0.13	81.78 ± 0.10	70.38 ± 0.09	84.89 ± 0.11
CE-Net	92.12 ± 0.09	84.89 ± 0.10	79.32 ± 0.09	89.44 ± 0.08
DANet	94.14 ± 0.08	86.56 ± 0.09	80.15 ± 0.07	91.15 ± 0.08
D2ANet	$\mathbf{99.78 \pm 0.06}$	$\mathbf{91.64 \pm 0.07}$	$\mathbf{85.11 \pm 0.04}$	$\mathbf{96.79 \pm 0.06}$

Table 2: Comparison with state-of-the-art on CRL dataset.

Methods	Acc	DSC	JSC	Sen
UNet	85.21 ± 0.11	75.78 ± 0.09	63.52 ± 0.12	81.69 ± 0.11
CE-Net	89.81 ± 0.08	79.48 ± 0.10	67.48 ± 0.08	85.59 ± 0.11
DANet	86.76 ± 0.10	77.26 ± 0.09	61.76 ± 0.08	90.37 ± 0.07
D2ANet	$\mathbf{91.84 \pm 0.05}$	$\mathbf{90.65 \pm 0.07}$	$\mathbf{83.32 \pm 0.05}$	$\mathbf{91.84 \pm 0.06}$

3.2. Training Settings and Metrics

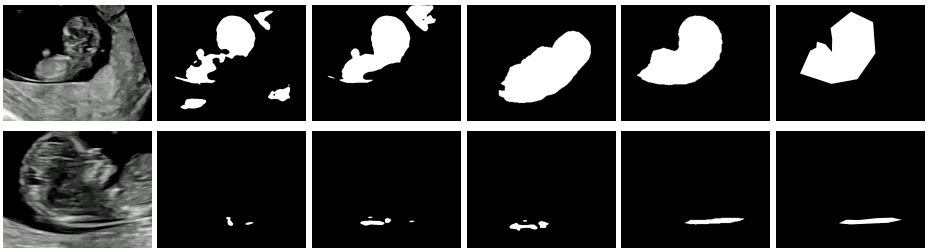
We implemented D2ANet using PyTorch. We use a pretrained ResNet34 as a backbone for UNet, CENet and D2ANet. The models were trained for 200 epochs with a batch size of 16. We use an Adam optimizer with the default initial learning rate of 3.10^{-3} and weight decay of 10^{-4}. We use the poly learning rate policy by multiplying the initial rate with $(1 - epoch/maxEpochs)^{0.9}$. Models are trained on a 12 GB TitanX GPU. To evaluate the models' performance, we used the following established metrics: accuracy (Acc), Jaccard similarity (JSC), Dice score (DSC), and sensitivity (Sen). We use the Dice coefficient loss with regularization (weight decay).

3.3. Quantitative Results

For a fair comparison of our proposed D2ANet, we choose two categories of approaches: (1) models based on encoder-decoder structure (UNet and CENet) and (2) a model based on spatial and channel attentions (DANet). Table 1 summarizes a direct comparison with existing experimental results on NT segmentation dataset, with D2ANet results being averaged over four training episodes. The experiments show that we were able to outperform all the algorithms on all metrics. For the CRL segmentation, Table 2 summarizes the comparison results of our D2ANet model against the same algorithms. The experiments show that D2ANet outperforming previous models by a significant ($> 6\%$) margin when considering three metrics (Acc, DSC and JSC). These results illustrate our model's strong learning ability even from challenging US segmentation tasks such as CRL and NT.

3.4. Qualitative Results

It is always important to look at the qualitative results of the images when measuring the effectiveness of a segmentation model. We have selected different images from the test sets of both NT and CRL segmentation datasets. Fig. 5 offers a preview of the raw data, followed by the predictions of UNet, CENet, DANet and D2ANet and the corresponding GT mask. For NT segmentation, it is clear that our model



(a) US (b) UNet (c) CENet (d) DANet (e) Ours (f) GT

Fig. 5: Qualitative performance. The columns show the US input frames (CRL on top and NT on the bottom), segmentation predictions of UNet, CENet, and DANet against our proposed method followed by the ground truth (GT), respectively.

Table 3: Ablation study of the D2ANet on NT dataset.

Methods	Acc	DSC	JSC	Sen
Model 1	94.64 ± 0.07	85.96 ± 0.08	80.11 ± 0.06	90.98 ± 0.07
Model 2	98.12 ± 0.06	88.13 ± 0.05	82.17 ± 0.07	94.12 ± 0.05
Model 3	97.87 ± 0.06	87.56 ± 0.05	81.35 ± 0.06	92.78 ± 0.05
D2ANet	$\mathbf{99.78 \pm 0.06}$	$\mathbf{91.64 \pm 0.07}$	$\mathbf{85.11 \pm 0.04}$	$\mathbf{96.79 \pm 0.06}$

can precisely locate and segment the area of tissue at the back of the fetus neck. We notice a slightly better performance of DANet compared to UNet and CENet, which failed in their segmentation. The same remark is valid in CRL segmentation, where our proposed D2ANet shows better segmentation performance compared to the other algorithms.

3.5. Ablation Study

We implemented a set of ablation studies to verify the effectiveness of D2ANet components, the contribution of spatio-channel fusion block, and the impact of dense connectivity. All the experiments were done on the NT segmentation dataset. In summary, the following models are compared:

- **Model 1**: the baseline model, the network without dense connectivity and SCF block
- **Model 2**: baseline model with the SCF block, and
- **Model 3**: baseline model while adding dense connectivity

In all the previous models, we replace the SCF block with a simple adding operation. The results are reported in Table 3. The addition of each component, including SCF block and dense connectivity, each contributes a considerable performance increase over the baseline on all segmentation metrics. The SCF block has the highest positive effect on performance metrics.

4. CONCLUSION

We have presented a novel segmentation network (D2ANet) tailored to predict NT and CRL structures from first trimester fetal US video scans. Our network differs from the previous methods based either on encoder-decoder structure or dual parallel/sequential attentions due to its architecture design and structure which combines two main ideas: SCF block and dense connectivity in a novel way leading to better segmentation. The experiments on both NT and CRL datasets, show that our model outperforms all the benchmark approaches on all segmentation metrics. The future work will study the performance of D2ANet on other medical image modalities.

5. REFERENCES

[1] L Drukker et al., "Vp18. 07: First trimester scans: how much time does it take to acquire the crl and nt?," *UOG*, vol. 58, pp. 174–174, 2021.

[2] P Taipale et al., "Learning curve in ultrasonographic screening for selected fetal structural anomalies in early pregnancy," *Obstetrics & Gynecology*, 2003.

[3] S Nirmala et al., "Measurement of nuchal translucency thickness in first trimester ultrasound fetal images for detection of chromosomal abnormalities," in *Proc. INCACEC*. IEEE, 2009, pp. 1–5.

[4] Z Sobhaninia et al., "Fetal ultrasound image segmentation for measuring biometric parameters using multi-task deep learning," in *Proc. IEEE EMBC*. IEEE, 2019.

[5] O Ronneberger and other, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[6] Z Gu et al., "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE TMI*, 2019.

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

[8] S Rueda et al., "Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: a grand challenge," *IEEE TMI*, vol. 33, no. 4, pp. 797–813, 2013.

[9] LH Lee et al., "Calibrated bayesian neural networks to estimate gestational age and its uncertainty on fetal brain ultrasound images," in *ASMUS*. Springer, 2020.

[10] R Yasrab et al., "End-to-end first trimester fetal ultrasound video automated crl and nt segmentation," in *IEEE ISBI*. IEEE, 2022, pp. 1–5.

[11] Jie Hu et al., "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF*, 2018, pp. 7132–7141.

[12] J Fu et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF*, 2019.

[13] "Fetal anomalie screen programme handbook," Report, NHS Screening Programmes, London, UK, 2015.