COMFormer: Classification of Maternal-Fetal and Brain Anatomy using a Residual Cross-Covariance Attention Guided Transformer in Ultrasound

Md Mostafa Kamal Sarker*, Vivek Kumar Singh*, Mohammad Alsharid, Netzahualcoyotl Hernandez-Cruz, Aris T. Papageorghiou, J. Alison Noble.

Abstract*—Monitoring the healthy development of a fetus requires accurate and timely identification of different maternal-fetal structures as they grow. To facilitate this objective in an automated fashion, we propose a deep-learning-based image classification architecture called the *COMFormer* to classify maternal-fetal and brain anatomical structures present in two-dimensional fetal ultrasound images. The proposed architecture classifies the two subcategories separately: maternal-fetal (abdomen, brain, femur, thorax, mother's cervix, and others) and brain anatomical structures (trans-thalamic, trans-cerebellum, trans-ventricular, and non-brain). Our proposed architecture relies on a transformer-based approach that leverages spatial and global features by using a newly designed residual cross-variance attention (R-XCA) block. This block introduces an advanced cross-covariance attention mechanism to capture a long-range representation from the input using spatial (e.g., shape, texture, intensity) and global features. To build *COMFormer*, we used a large publicly available dataset (BCNatal) consisting of $12,400$ images from 1,792 subjects. Experimental results prove that *COMFormer* outperforms the recent CNN and transformer-based models by achieving 95.64% and 96.33% classification accuracy on maternal-fetal and brain anatomy, respectively.

***Index Terms*—Fetal ultrasound, maternal-fetal, deep learning, convolutional neural network, transformer.**

I. Introduction

PRENATAL ultrasonography is a non-invasive, real-time imaging modality employed during pregnancy. Ultrasound (US) scanning has the advantage of using non-ionizing radiation, is convenient, and is safer for pregnant women than other imaging modalities like magnetic resonance imaging (MRI) or computed tomography (CT). In two-dimensional US, fetal biometry, consisting of measurement of the head circumference (HC), femur length (FL), biparietal diameter (BPD), and abdominal circumference (AC), is used for gestational age estimation and fetal growth monitoring [1]. US can also assist the functional health of the fetus, such as the assessment of the fetal heart rate and the bladder.

Acquisition of fetal US standard planes and anatomical structure classification has received significant attention in the recent literature [2]–[4]. Bridge *et al.* [5] identified the fetal heart from every frame of the US scan videos by applying a particle-filtering-based method. Chen *et al.* [6] use domain-transferred CNNs to classify fetal abdomen images from non-abdomen images with parameter weights transferred from a CNN trained on natural images. Yaqub *et al.* [7] presented a random forest-based method for classifying six fetal US planes. In related work, Carneiro *et al.* used a probabilistic boosting-tree to detect and classify fetal anatomical structures [8]. In another work, Yaqub *et al.* [9] also developed a system that investigates if all relevant anatomical fetal views have been included in a subject's imaging record. Baumgartner *et al.* [10] proposed SonoNet, a pre-trained CNN architecture focused on detecting 13 planes from US video clips using a supervision-based CNN method. Savioli *et al.* [11] investigated the automatic measurement of the vascular diameter of the fetal abdominal aorta from US images using a CNN with a convolution-gated recurrent unit (C-GRU). The C-GRU takes advantage of the signal's temporal redundancy, and a regularised loss function called CyclicLoss to improve prior knowledge about the periodicity of the observed signal. Yasrab *et al.* [12] and Chen *et al.* [13] use spatio-temporal methods to classify fetal US images at the frame level.

There is also work exploring using augmentation-based approaches or auxiliary information to improve the performance of CNN architectures for standard plane detection. Specifically, Lee *et al.* [14] introduces a particular approach to augmentation to improve the performance of fetal standard plane classification models. Ahmed *et al.* [15] and Cai *et al.* [16] make use of visual heatmaps to assist in identifying standard planes depicting the abdominal circumference in fetal US images. A multi-scale model integrated with an attention mechanism is used by Xhi *et al.* [17] where the author segments the fetal heart and lungs in US images. A recent study with 12,400 US images from 1,792 subjects [18] classifies the common maternal-fetal US images using some recently made available ImageNet pre-trained CNN models. That study [18] makes their dataset publicly available and is used in this paper to enable benchmarking of our approach.

Transformer-based methods have achieved great attention and some success in the medical imaging domain [19]. As an alternative to CNNs, transformers can extract long-range dependencies and highlight prominent feature representations through their self-attention mechanism. A limited study has reported the effectiveness of recent transformers-based ap-

M.M.K.S, M.A., N.H., and J.A.N are with the Institute of Biomedical Engineering, University of Oxford, Oxford, OX3-7DQ, UK (corresponding author e-mail: md.sarker@eng.ox.ac.uk).
A.T.P is with the Nuffield Department of Women's Reproductive Health, University of Oxford, Oxford, UK.
M.A. is also with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE.
V.K.S is with the School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK.

*Contributed equally to this work.

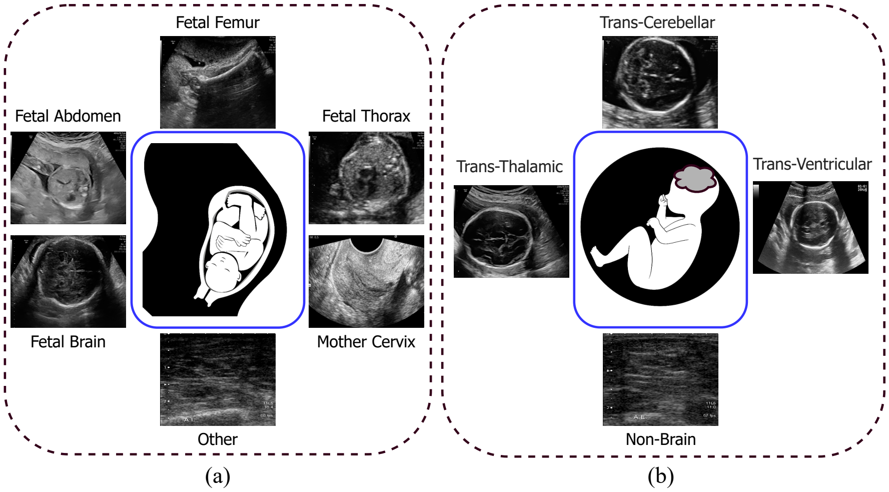


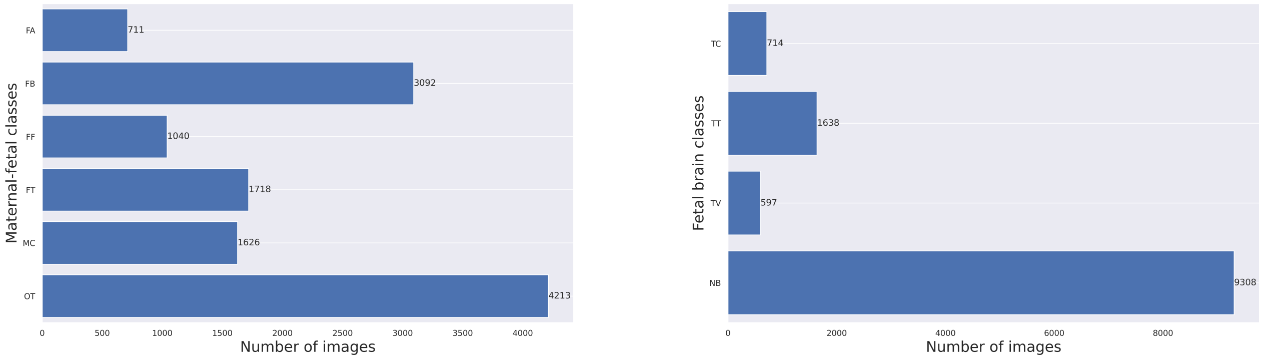
Fig. 1: Example images from the "BCNatal" dataset [18] (a) for the maternal-fetal anatomy dataset, specifically, the abdomen, brain, femur, thorax, mother's cervix, and any "other" US plane, and (b) for the brain anatomy dataset, specifically, the trans-thalamic, trans-cerebellum, trans-ventricular and "non-brain" US plane.

proaches in US image analysis. Gheflati *et al.* [20] used a vision transformer and different augmentation strategies to classify breast US images. Plotka *et al.* [21] introduced a model called BabyNet that predicts the birth weight of the fetus from an US video. Yang *et al.* [22] propose a fetal head circumference auto-measurement method that combines a transformer and a CNN to extract a meaningful feature representation that incorporates both the local and global features from the US images.

1) Motivation: Detecting anatomical planes for fetal assessment is recognized as a highly-skilled task in which the quality of ultrasound decision-making depends on a sonographer's skill. Even when a sonographer is trained to a high level, there can be significant inter- and intra-observer acquisition variability [23]–[26]. The motivation behind this work is to support sonographers in detecting standardised planes promptly, allowing for more optimal use of expert sonographers' and operators' time without compromising the quality and accuracy of maternal-fetal structure identifications. The need for such support is more apparent when considering scenarios requiring the determining of different views of the same structure or when classifying substructures within the same anatomical structure. Due to the quality of its prediction and the speed of the automated solution compared to manual identification, *COMFormers* are well-suited for implementation in a real-life clinical setting. Automating the clinical workflow may allow more subjects to be screened in a clinic session. Classifying anatomical structures - maternal or fetal - through COMFormers may support that objective.

2) Contribution: This paper presents a new automated vision transformer-based architecture called *COMFormer*, which uses a Residual Cross-Covariance Attention Guided Transformer to classify maternal-fetal and brain anatomical structures in fetal ultrasound images. The proposed architecture is inspired by the cross-covariance attention (XCA) block from previous work [27], which enhances the discriminability of key features in targeted regions. The XCA plays a critical role in determining the relevance of input tokens to others. However, this process can be computationally expensive and lead to small gradients in deeper networks. The proposed residual cross-variance attention (R-XCA) block utilises residual connections to prevent minimal gradients and facilitate the learning process of *COMFormer*. These connections also play a crucial role in preserving the original spatial information from the input, which is necessary and important for fetal US image classification. With the R-XCA block, our model can effectively focus on different parts of the fetal input image while residual connections ensure that the original spatial details are not lost. The *COMFormer* integrates low and high-level information such as texture, shape, and anatomy boundaries to improve feature extraction. Additionally, it includes local patch interaction (LPI), feed-forward network (FFN), and fully connected (FC) layers to improve information flow between channels and extract per-patch information. We evaluate the performance of the *COMFormer* architecture on a publicly available fetal ultrasound dataset of $1,792$ subjects. The proposed model is used to classify maternal-fetal anatomy into six different anatomical planes (abdomen, brain, femur, thorax, mother's cervix, and others) and fetal brain anatomy (trans-thalamic, trans-cerebellum, trans-ventricular, and non-brain) into four sub-classes. It is important to note that the *COMFormer* architecture is used separately to train and evaluate the above two classification tasks. Our experiments demonstrate that the *COMFormer* architecture achieves competitive classification results compared to recent CNN and transformer-based methods. We also conduct a design justification study to validate the effectiveness of each component in the *COMFormer* architecture.

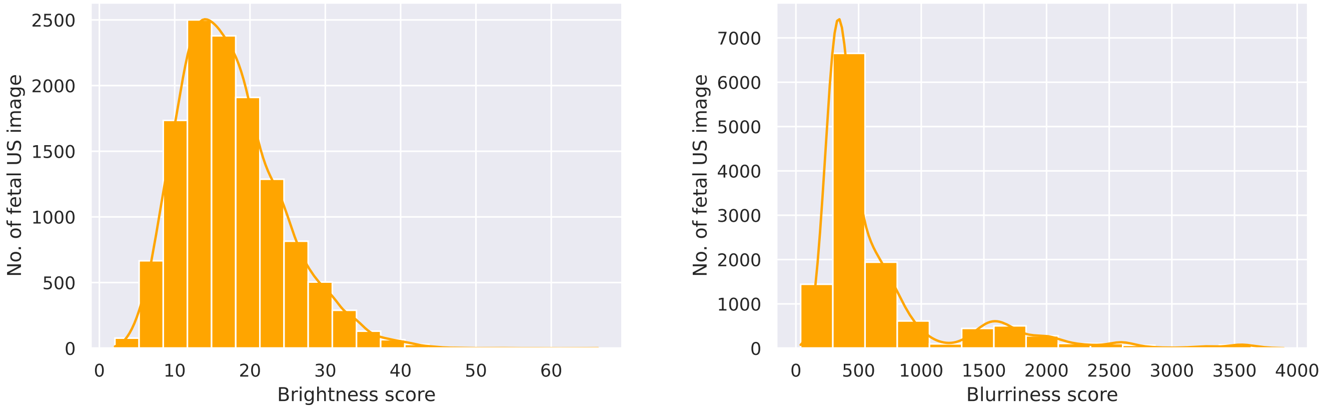
The paper is structured as follows. Section II provides an overview of the "BCNatal" dataset and introduces the proposed *COMFormer* architecture. Section III presents the experimental results, comparing various CNN and transformer-based models. Finally, Section IV concludes the paper by summarizing the main findings and suggesting possible directions



(a)

(b)

Fig. 2: Class-wise image distribution of “BCNatal”dataset [18], (a) for the maternal-fetal anatomy dataset, including fetal abdomen (FA), brain (FB), femur (FF), and thorax (FT), mother’s cervix (MC) and any other (OT) US plane, (b) for the brain anatomy dataset, including trans-thalamic (TT), trans-cerebellum (TC), trans-ventricular (TV) and rest of the non-brain (NB) US plane.



(a)

(b)

Fig. 3: Histograms of image quality metrics computed for the 12,400 fetal US images: (a) brightness score and (b) blurriness score.

for future research.

II. Materials and Method

This section provides an overview of the fetal US dataset used in the study and describes the architectural design of a Residual Cross-Covariance Attention (R-XCA) block, along with its interconnected elements, for the identification of fetal anatomical structures in US images.

A. Dataset

We used the publicly available “BCNatal” dataset [18]. It consists of $12,400$ fetal US images from $1,792$ subjects and was collected between October 2018 and April 2019 from the routine clinical procedures during the second and third trimesters of the pregnancy screening. Fig. 1 shows some example images from the “BCNatal” dataset. The fetal gestational age ranged from 18 to 40 weeks based on crown-rump length measurement. Several operators collected all the fetal US images with similar experiences from a total of six different US machines, including three Voluson E6 (GE Medical Systems, Zipf, Austria), one Voluson S8, one Voluson S10, and one Aloka (Aloka CO., LTD.) by using an abdominal curved probe with a frequency range from 3 to 7.5MHz and a 2 to 10-MHz vaginal transducer. The original dataset is split into two sub-sets: the maternal-fetal dataset, including fetal anatomy abdomen (FA), brain (FB), femur (FF), thorax (FT), mother’s cervix (MC), and any other (OT) US plane, and the brain anatomy dataset, including trans-thalamic (TT), trans-cerebellum (TC), trans-ventricular (TV), and rest of the other non-brain(NB) US plane. A single expert manually labelled all images. Note that all the images are stored in ‘.png’ format and contain no patient metadata information. Fig. 2 illustrates the class-wise image distributions of the datasets.

1) Image Quality Assurance: To ensure the best US image quality before feeding a training image into the deep neural network, we used two image quality metrics called *bluriness* scores [28], [29] and *brightness*. **Blurriness score**: The blurriness score was estimated by utilizing the variance of the intensity of the fetal ultrasound (FUS) image, denoted as $I_{FUS}(X_c, Y_c)$, which was smoothed using a Gaussian filter $G_f(X_c, Y_c)$, as described in [29]. The Gaussian filter is mathematically expressed as follows:

$$G_f(X_c, Y_c) = \frac{1}{(2\pi\sigma^2)} e^{-\frac{\left(X_c{}^2 + Y_c{}^2\right)}{2\sigma^2}}, \tag{1}$$

Here σ represents the standard deviation of the Gaussian distribution, and X_c and Y_c denote the image coordinate of $I_{FUS}(X_c, Y_c)$. Additionally, we employed the Laplacian operator to calculate the gradient variation (∇I_{FUS}) of the image I_{FUS} in two dimensions. This was achieved by summing the second partial derivatives in Cartesian coordinates, yielding the following expression:

$$\nabla^2 I_{FUS}(X_c, Y_c) = \frac{\partial^2 I_{FUS}}{\partial X_c{}^2} + \frac{\partial^2 I_{FUS}}{\partial Y_c{}^2}, \tag{2}$$

It is worth noting that a low score indicated the image was blurry, while a high value showed that the fetal ultrasound image was sharp based on the measured gradient variation. **Brightness score**: It assists in identifying various fetal ultrasound image characteristics. Sometimes, the acoustic shadows or excess of subcutaneous fat present in the patient absorb the energy that creates a poor image. In this regard, we employed the brightness estimation algorithm proposed by [30]. Fig. 3 plots a histogram of the US image quality scores for all image samples. In training and testing, we ignored the images with a brightness score lower than 20 (see Fig. 3a), which corresponds to a very dark anatomical region in the US image. An image with a blurriness score of less than 100 (see Fig. 3b) was also not used in training or testing. Overall, we exclude the 615 (5%) number of the images of the entire dataset.

B. COMFormer Architecture

Our *COMFormer* architecture incorporated a residual cross-covariance attention block (R-XCA) block to leverage both local (i.e., shape, texture, intensity, etc.) and global fetal US image features using a cross-covariance attention mechanism to learn learning long-range representation. As shown in Fig. 4, the *COMFormer* incorporated three main blocks: R-XCA, local patch interaction (LPI), and feed-forward network (FFN). We provide a brief description of each layer below.

1) Residual Cross-Covariance Attention Block (R-XCA): The proposed R-XCA is an attention technique highlighting the meaningful features from input US. The R-XCA functions between the channel's dimension instead token in the input sequence. In other words, it minimizes the quadratic complexity between each dimension of the token embeddings. It is important to note that the token represents the whole US image when each image is divided into a series of patches that feed into the *COMFormer* input.

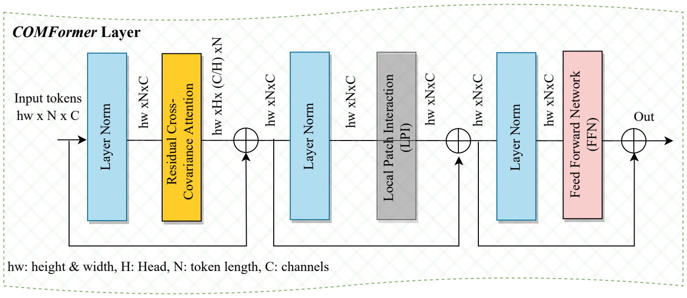

COMFormer Layer
Input tokens
hw x N x C
Layer Norm
hw xNxC
Residual Cross-Covariance Attention
hw xHx (C/H) xN
hw xNxC
Layer Norm
hw xNxC
Local Patch Interaction (LPI)
hw xNxC
hw xNxC
Layer Norm
hw xNxC
Feed Forward Network (FFN)
Out
hw: height & width, H: Head, N: token length, C: channels

Fig. 4: Overview of the *COMFormer* layer.

To apply the self-attention mechanism, consider a sequence of N_T entities (I_1, I_2,.....,I_n) by $I \in \mathbb{R}^{N_T \times d_i}$ where N_T corresponds to the number of tokens, and each dimensionality d_i. Note that d_i includes the $batch$, $token$, and d_m refers to the batch size, the number of elements for sequence, and the dimensions of the embedding vector per US input element to the sequence, respectively. Input image I is linearly projected to queries (Q), keys (K), and values (V). Note that queries are a set of vectors that need to estimate the attention, and keys are also a set of vectors demanded to calculate the attention against others. The queries, keys, and values leveraged the weight metrics of $W_q \in \mathbb{R}^{d_i \times d_q}$, $W_k \in \mathbb{R}^{d_i \times d_k}$, and $W_v \in \mathbb{R}^{d_i \times d_v}$ that extract the feature representation. The output of Q, K and V measured the $K = IW_k$, $Q = IW_q$, and $V = IW_v$ respectively.

The main aim of the self-attention mechanism is to capture the relation between all of the N_T individuals through encoding entities concerning global contextual information. Therefore, the attention maps can be expressed as follows [27]:

$$\mathcal{A}(K, Q) = \text{Softmax}\left(QK^\top / \sqrt{d_k}\right) \tag{3}$$

Where Q, K, and V stand for the queries, keys, and values, respectively. The dimension of the queries and keys is denoted by d_k. Furthermore, the self-attention outcome is a weighted accumulation of the T_N token features in V, where the weights are formulated as:

$$\text{Attention}(Q, K, V) = \mathcal{A}(K, Q)V \tag{4}$$

The cross-covariance attention [27] estimate the covariance among the features of the key and query matrices can be defined as:

$$\begin{aligned} \text{XC-Attn}(Q, K, V) &= V\mathcal{A}_{\text{IC}}(K, Q), & (5) \\ where, \mathcal{A}_{\text{IC}}(K, Q) &= \text{Softmax}\left(\hat{K}^\top \hat{Q} / \tau\right) & (6) \end{aligned}$$

Here the $Softmax$ is used to generate the attention vectors, and τ is the learnable temperature that allows the convergence of the network training.

Finally, the residual cross-covariance attention (R-XCA) block can be written as:

$$\text{R-XCA}(Q, K, V) = \text{XC-Attn}(Q, K, V) + I \tag{7}$$

The R-XCA model represents an attention-based approach to leverage the cross-covariance of patch representation projections (keys and queries) at each layer and automatically constructs one-dimensional filters. These filters are then used to extract relevant meaningfulness information from every patch.

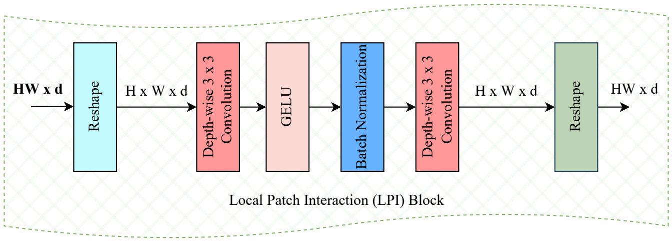

HW x d
Reshape
H x W x d
Depth-wise 3 x 3 Convolution
GELU
Batch Normalization
Depth-wise 3 x 3 Convolution
H x W x d
Reshape
HW x d
Local Patch Interaction (LPI) Block

Fig. 5: Architectural details of the Local Patch Interaction (LPI) block.

2) Local Patch Interaction (LPI) Block: Fig. 5 shows the detailed architectural description of the LPI block. It belongs to the standard convolutional block that utilizes some tensor reshaping and permuting procedures. It involves two depth-wise 3×3 convolutional layers with in-between batch normalization and a non-linear Gaussian Error Linear Unit (GELU) [31] activation function. The LPI block added directly after each R-XCA block provides communication between patches. In other words, it provides a solution to combine knowledge between tokens in the US input sequence. The first convolutional layer has a kernel size of 3×3. The second convolutional layer follows the first layer structure except for the output, which relies on the number of input channels. The layer normalization provides the normalization of all the activations of an individual CNN layer from a batch by accumulating statistics from a single training case. Usually, the output of the R-XCA block keeps the shape of $HW \times d$. Note that H, W , and d correspond to the height, width, and depth, respectively. However, performing a standard convolution operation in a 2-dimensional plane requires reshaping the size to $H \times W \times d$. Once the features are processed, it is reshaped to the original shape of $HW \times d$.

3) Feed-Forward Network (FFN): The Feed-Forward Network (FFN) consists of a single hidden layer that incorporates the four-dimensional hidden components. For an input I, FNN can be expressed as follows:

$$\mathrm{FFN}(I) = \sigma\left(IW_1 + b_1\right)W_2 + b_2 \tag{8}$$

Here, σ is a GELU activation function. FFN encourages interaction across all features when there is no feature interaction in the LPI block.

4) Class Attention: We also employed class attention layers that average the patch embeddings of the previous *COMFormer* layer by assigning them to a class (CLS) token via single-way attention between the CLS tokens and the patch embeddings. Note that the class token attends to the patches and provides the most suitable regions of a US image.

C. Cost Function

Since the dataset [18] is imbalanced, choosing an appropriate loss function is critical for successfully training a deep learning-based model. Our study used two cost functions: Cross-Entropy (CE) and Focal Loss (FL). The CE loss is defined as:

$$\mathcal{L}^{\mathrm{CE}}(fe_{gt}, fe_{pd}) = -\sum_{j=1}^{m} fe_{gt}.log(fe_{pd_j}) \tag{9}$$

where m refers to the number of classes (in our case, six and four classes), fe_{gt} is the label defined by the clinical expert (ground-truth), and fe_{pd_j} is the softmax probability of the j^{th} class.

The FL [32] applies a scaling factor to the softmax CE loss to reduce the associated loss for successfully identified instances while concentrating on challenging ones. FL is defined as:

$$\mathcal{L}^{\mathrm{FL}}(fe_{gt}, fe_{pd}) = -\sum_{j=1}^{m} (1 - fe_{pd_j})^{\gamma}.log(fe_{pd_j}) \tag{10}$$

where the hyperparameter $\gamma \in [0.5, 2]$.

D. Training Details

We employed the same dataset distribution as stated in previously published work [18]. We rescaled the original input fetal US image to a spatial resolution of 224×224 pixels. To avoid overfitting, we applied data augmentation, specifically rotation of 30 degrees, horizontal and vertical flipping with a probability of 0.5, and a scaling factor of 0.25. We computed the mean and standard deviation of intensity across a single-channel US image and normalized all the samples accordingly. The model had a patch size of 16×16, embedded dimension of 768, depth, and a number of heads was each set to 12. We used an SGD optimizer with an initial learning rate of 0.001. All models were trained for 100 epochs with a batch size of four. At the end of the training, all model weights were saved based on the highest classification accuracy yielded on the validation set. Note that we trained all the methods employed with their default hyperparameter settings proposed in their literature. We used the same input size, loss function, and number of epochs to train the proposed COMFormer and compared methods. We report classification results using four established performance evaluation metrics: accuracy, precision, recall, and F1-score.

III. Experimental Results

This section assesses the effectiveness of the proposed *COMFormer* and compares it against existing CNN and transformer-based models that were pre-trained on ImageNet. We examine various model versions with different input resolutions and investigate the impact of data augmentation and loss function type. Initially, we conducted experiments for the six-class maternal-fetal anatomy classification and then applied the best-performing *COMFormer* configuration to the four-class brain anatomy classification task.

A. Comparison of CNN and Transformers-based Models to Classify the Maternal-fetal Anatomical Plane

Table I presents a comparison results between the proposed *COMFormer* architecture and previously published models for maternal-fetal anatomical plane classification. To make a fair

TABLE I: Comparison of the proposed model performance (%) for classifying the maternal-fetal anatomical plane with the four CNN and five recent transformers-based models. The best significant scores are in bold.

Methods	**Accuracy**	**Precision**	**Recall**	**F1-score**
CNN-based				
VGG-16 [33]	92.28	90.78	92.02	91.39
ResNet-101 [34]	93.06	91.53	93.59	92.54
ResNetXt-101 [35]	93.37	91.87	94.08	92.96
DenseNet-169 [36]	93.50	92.51	93.88	93.18
Transformers-based				
ViT [37]	93.29	92.13	93.49	92.80
BEiT [38]	93.93	93.14	93.91	93.52
CaiT [39]	93.41	93.58	93.99	93.78
Swin [40]	93.47	93.51	93.60	93.55
XCiT [41]	93.59	93.62	94.03	93.82
COMFormer (w/o QA)	94.76	93.12	94.55	93.83
COMFormer	**95.64**	**94.65**	**95.87**	**95.23**

comparison, we fine-tuned four CNN-based models pre-trained on ImageNet from; VGG-16 [33], ResNet-101 [34], ResNetXt-101 [35], and DenseNet-169 [36]. Similarly, we also fine-tuned five recent vision transformer-based models; ViT [37], BEiT [38], CaiT [39], Swin Transformer [40], and XCiT [27]. Experimental results confirmed that *COMFormer* achieved the highest classification results with the accuracy score of 95.64% than the other methods.

In CNN-based models, we found that DenseNet-169 marginally yielded the best results of 93.50% with the second-highest ResNeXt-101. Each DenseNet-169 layer received collective US feature information from the previous layers using a densely connected CNN. In simple terms, the last output layer collects all the features representation from every single layer, which contributes to better classification performance than for other compared methods. VGG-16 and ResNet-101 obtained similar classification performance in all the evaluation metrics. For the transformer-based models, BEiT achieved the second-best results to *COMFormer* with an approximate 2% for all the measured metrics. ViT, CaiT, and Swin performed similarly by achieving results in the range of 93%. These methods process the input US into small patches through a self-attention mechanism that only emphasizes capturing global features of maternal-fetal anatomical plane structure features and has limited aggregation of local spatial information. Additionally, we have performed experiments without the quality assurance

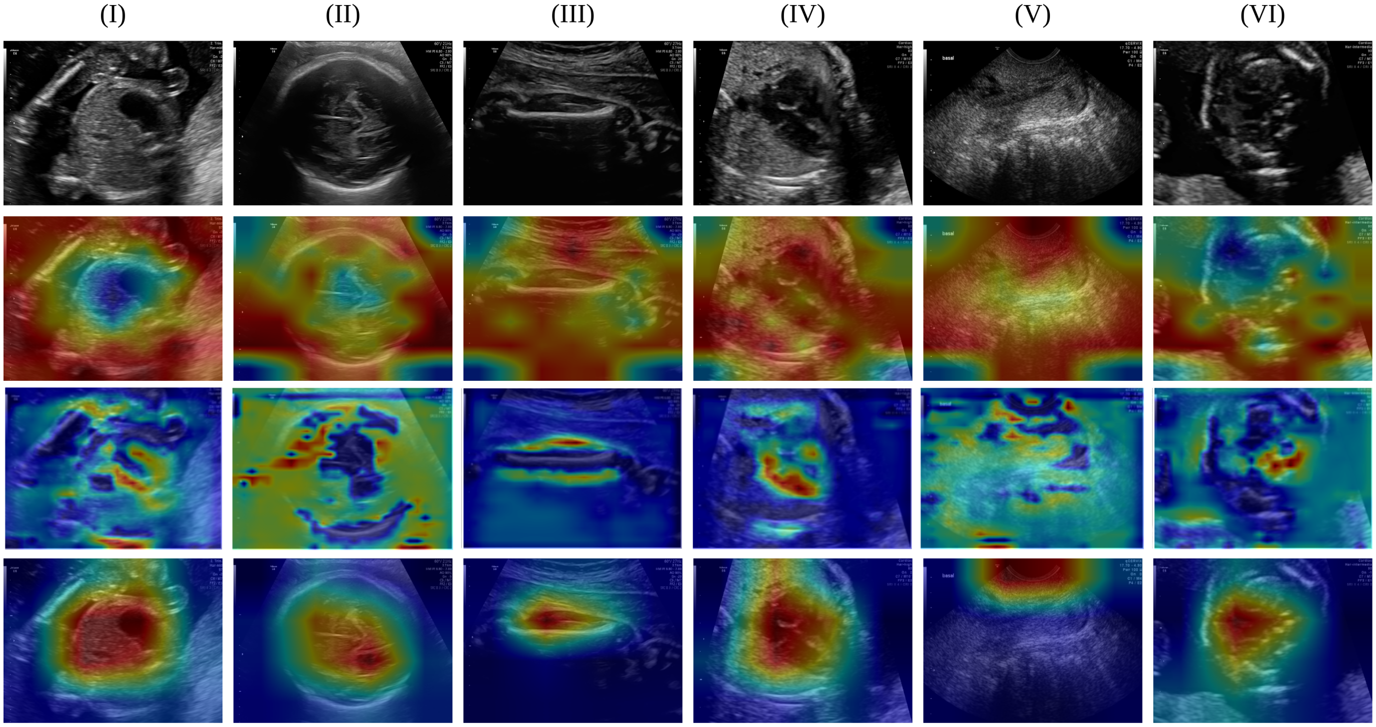


Fig. 6: Visualisation results of the activation maps. For every column, we show an input maternal-fetal anatomical plane, the corresponding activation maps from the outputs of the CNN-based DenseNet-169, XCiT, and the *COMFormer* model. Note that the input image of the first row from left to right is the fetal abdomen (FA), fetal brain (FB), fetal femur (FF) and fetal thorax (FT), mother's cervix (MC), and any other (OT) image plane, respectively. The proposed model has captured the most important anatomical structures in the presence of several imaging artefacts. The residual attention highlighted the salient features and ignored the unwanted ones.

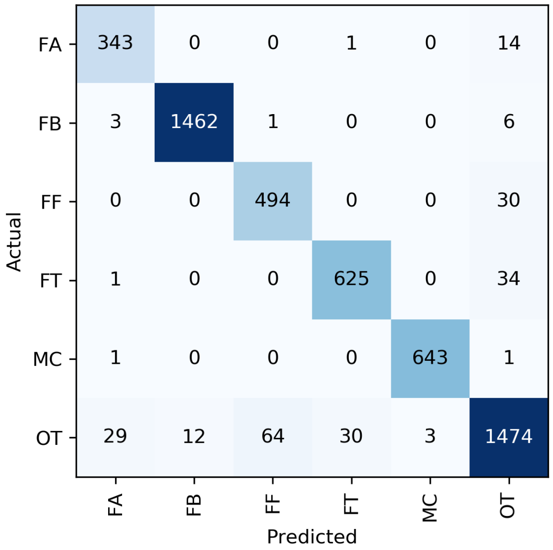
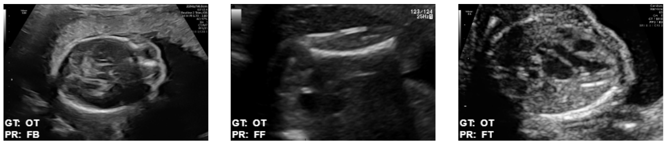


Fig. 7: Illustration of a confusion matrix using *COMFormer* for classifying the maternal-fetal anatomical plane in the US.



(a) (b) (c)

Fig. 8: Misclassification of the OT class images predicted by the *COMFormer*. Note that (a), (b), and (c) refers to the FB, FF, and FT classes, respectively. Here, GT and PR belong to the ground-truth and prediction, respectively.

(w/o QA) step when training the COMFormer and found that it decreased the classification results with 1% when compared against utilizing it.

To complement our quantitative findings, Fig. 6 shows some examples of activation maps of the best performance achieved by the DenseNet-169, XCiT, and *COMFormer* architecture. It is important to note that the intensity of the red colour represents where the model pays more attention when making the final prediction, and blue represents the less informative regions or pixels. From the visual inspection, the example images contain the six maternal-fetal anatomy structures with various imaging artefacts, including shadows, speckle noise, and neighbouring tissues. The DenseNet-169 highlighted most of the pixels present in the images. Its convolutional filters receptive field activated to capture the additional discussed artefacts. It also highlighted the hypoechoic tissues of the fetal abdomen, fetal brain, fetal femur, and other planes in examples I, II, III, and VI, respectively. However, examples IV and V refer to the FT and MC in which DenseNet-169 filters emphasize the neighbouring grey pixels that contain the higher speckle noise rather than focusing on the anatomy.

In turn, the XCiT model captured the anatomical structure but was not accurate as the proposed *COMFormer*. It paid more attention to the background regions and less to targeted pixels. The *COMFormer* model precisely captured the anatomical structures while ignoring the US imaging artefacts. The R-XCA block with a cross-covariance attention mechanism focused on the targeted anatomical pixels shown in all six examples.

TABLE II: The *COMFormer* model class-wise results for classifying the maternal-fetal anatomy in US.

Classes	Accuracy	Precision	Recall	F1-score
FA	95.81	90.98	95.81	93.33
FB	99.32	99.19	99.32	99.25
FF	94.27	88.37	94.27	91.23
FT	94.69	95.27	94.70	94.98
MC	99.68	99.54	99.69	99.61
OT	91.43	94.55	91.44	92.97

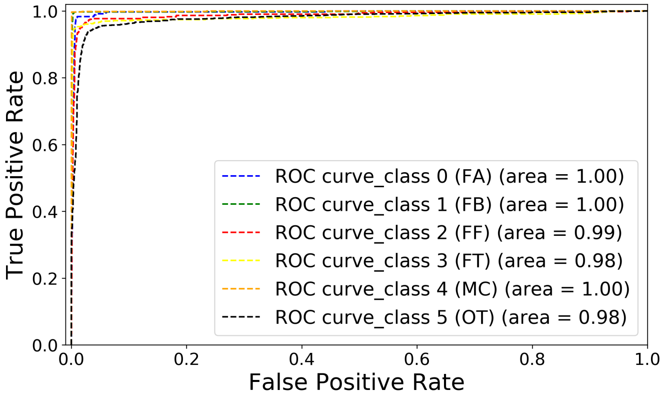


Fig. 9: The receiver operating characteristic (ROC) curve using *COMFormer* for the maternal-fetal anatomical plane classification task.

Fig. 7 shows the confusion matrix of the *COMFormer* for the maternal-fetal anatomical plane classification. From the experimental findings, the *COMFormer* correctly classified most of the maternal-fetal anatomical structures in the US, regardless of OT class, where 11% of the images were classified incorrectly into different categories. We discovered that some of the FT and FF class images are predicted as the OT. Similarly, the OT class images with 7.5% samples are confused with the FT and FF classes. The OT class features are mostly in-distributed with FT and FF classes, making it difficult for a model to differentiate precisely between them. We identified the misclassified samples labelled initially to the OT. With the help of *COMFormer* prediction scores and visual inspection of the misclassified cases, we found that they shared FF, FB, and FT classes images, as shown in Fig. 8. From the experimental findings and visual inspection, we found that the OT samples share similar features, as the *COMFormer* predicted them as their actual class based on their maternal-fetal anatomical structures.

Fig. 9 presents the ROC curve obtained using the *COMFormer* for the maternal-fetal anatomy plane classification task. All six classes achieved very high AUC with an average score of 99%. The *COMFormer* produced fewer misclassification and showed its strength in correctly predicting the multiple anatomical structures while ignoring the artefacts present in the US, as shown in Fig. 6. Moreover, the last

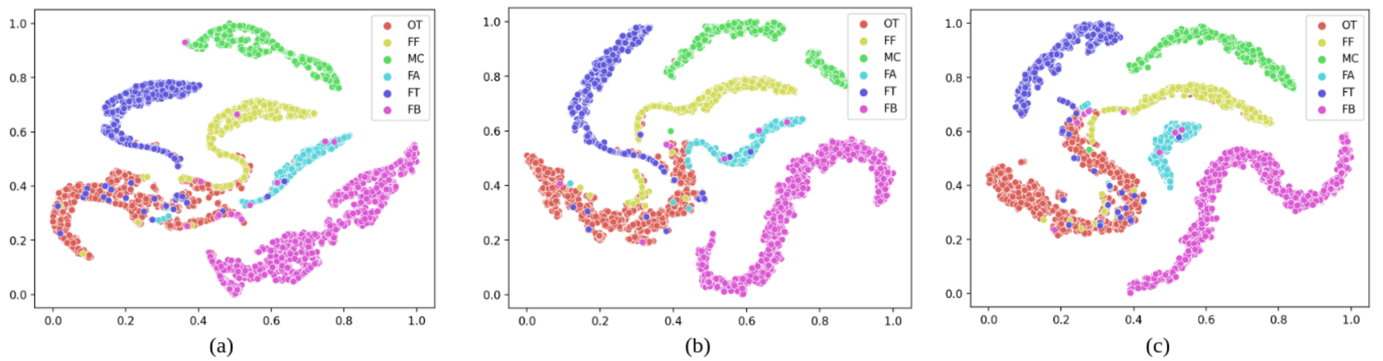


Fig. 10: t-SNE feature visualization of the maternal-fetal classification by (a) DenseNet-169, (b) XCiT, and (c) *COMFormer* .

layer features representations were obtained by embedding the 6-dimensional class vector of the DenseNet-169, XCiT, and *COMFormer* into 2 dimensions for maternal-fetal anatomy classification shown using t-distributed stochastic neighbour embedding (t-SNE) [42] in Fig. 10. The results indicate that DenseNet-169 and XCiT class feature clusters exhibited overlapping patterns with other classes, resulting in misclassification. We can graphically depict the inter-class variability of the classes. We can better analyze the heterogeneity among the test set samples using the t-SNE visualization. For example, FF, MC, FA, FT, and FB classes are distinguishable as a cluster. In contrast, we can identify classes with lower classification accuracy, like those overlapping with OT. In addition, Table II summarises the class-wise performance metrics of *COMFormer* with the accuracy score of 95.81%, 99.32%, 94.27%, 94.69%, 99.68%, and 91.43% to classify FA, FB, FF, FT, MC, and OT, respectively.

B. COMFormer Design Justification Experiments on Maternal-fetal Anatomy Classification

Next, we examine *COMFormer* performance by measuring the effects of changing input image resolution sizes, the impact of data augmentation, and loss functions on the performance of a six-class maternal-fetal anatomy classification task. Finally, we kept the best *COMFormer* parameters setting leveraged to train and evaluate in classifying the four-class fetal brain anatomy structures.

1) Effect of changing the input image resolution: In this study, we compared training with images of two different resolution sizes with 384×384 and 224×224 pixels. The dimensions of the original images are different, the heights and widths ranging from 787 to 196 and from 1605 to 280 pixels, respectively. Note that we resized the images using bilinear interpolation that contains their original resolution to the two discussed sizes. Table III shows the results for changing the input image size in classifying the maternal-fetal anatomy structures. We noticed that, with the resolution size of 224×224, the *COMFormer* achieved a higher performance of 1% than 384×384 for all the metrics. The fetal US contains additional noise that affects these grey pixels and does not help the model learn meaningful features. With the size of 224×224 pixels, holding less noise allows *COMFormer* that learns significant anatomical key spatial and global features in the US.

TABLE III: Evaluating the effect of variation in the input image size on *COMFormer* performance.

Image size	Accuracy	Precision	Recall	F1-score
384 ×384	95.33	94.51	95.14	94.82
224 ×224	**95.64**	**94.65**	**95.87**	**95.23**

2) Effect of data augmentation: Table IV presents the results of verifying the effect of adding data augmentation to *COMFormer*. As can be seen, adding the additional diversity in feature representation enhances classification results in the range of $0.5\% - 1\%$ in terms of accuracy, precision, recall, and F1 scores. The data augmentation fills the semantic gap and helps the *COMFormer* training to be more generalised.

TABLE IV: Evaluating the effect of with and without data augmentation using *COMFormer*.

Data aug.	Accuracy	Precision	Recall	F1-score
No	95.20	93.56	95.09	94.72
Yes	**95.64**	**94.65**	**95.87**	**95.23**

3) Effect of choice of the loss function: Table V compares models built with two different loss functions, CE and FL. Experimental results show that the model with FL outperforms the model with CE loss for all metrics with an improvement of $0.5 - 1\%$.

TABLE V: Evaluating the effect of the loss function using *COMFormer*.

Loss function	Accuracy	Precision	Recall	F1-score
CE	95.09	94.32	94.84	94.56
Focal	**95.64**	**94.65**	**95.87**	**95.23**

C. Brain Anatomy Classification

Table VI demonstrates the class-wise *COMFomer* performance for classifying the brain anatomical planes in the US. We already explored the best parameter of the *COMFomer* on the maternal-fetal classification task that translated to the brain anatomy classification task. We separately trained and evaluated the model for four-class brain anatomy prediction. The experimental results show *COMFomer* provided average

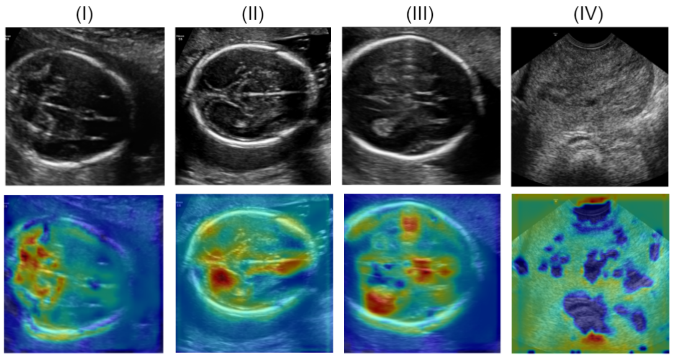


Fig. 11: Visualisation results of the activation maps for the brain anatomy class generated using the *COMFormer*. Each class presents a single example in which the model pays more attention to provide a final prediction. Examples I, II, III, and IV correspond to the TC, TT, TV, and NB.

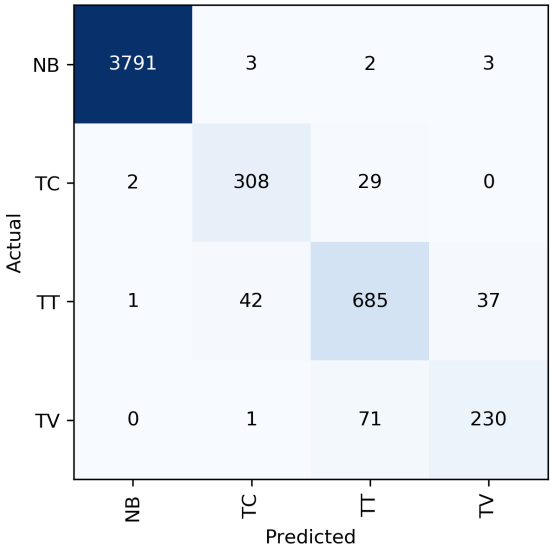


Fig. 12: Illustration of a confusion matrix using *COMFormer* for classifying brain anatomy plane.

accuracy and F1-scores of 89.08% and 89.37%, respectively. The model focused on capturing the TC, TT, and TV brain anatomy that ignored the neighbour background with artefacts present in the US, as shown in Fig. 11. However, the NB class contains an anatomical structure of not a brain presented in Fig. 11 (IV) where the model only highlights the majority of the grey pixels. In the remaining examples, such as I, II, and III, the *COMFomer* focuses on the targeted anatomical structures highlighted in red and ignores the background.

TABLE VI: The *COMFormer* model class-wise results for classifying the brain anatomy in US.

Classes	Accuracy	Precision	Recall	F1-score
NB	99.78	99.87	99.84	99.86
TC	90.86	90.54	84.66	87.50
TT	89.55	85.51	91.76	88.52
TV	76.15	86.62	77.15	81.61

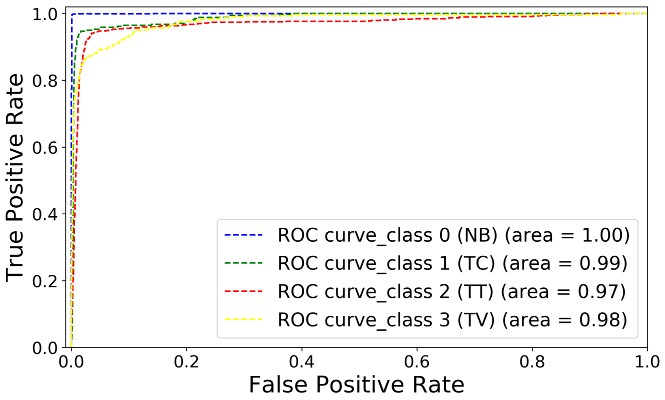


Fig. 13: Illustration of ROC curve using *COMFormer* to classify the brain anatomy in the fetal US.

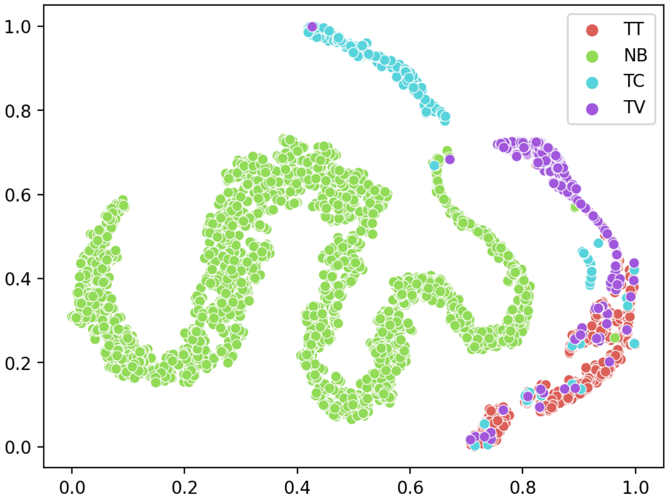
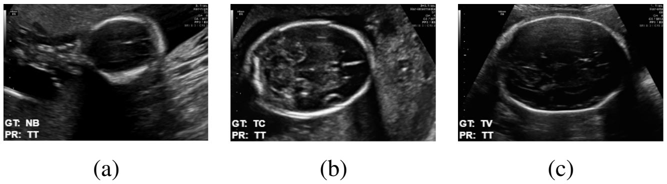


Fig. 14: t-SNE feature visualization of the brain anatomy classification by *COMFormer*.

Fig. 12 shows the confusion matrix of the brain anatomy planes classification obtained through the *COMFormer*. The NB contains the images which do not correspond to the brain structure classified correctly with only a 0.28% error rate. However, the TV has an error rate of 24% in which the images incorrectly predicted to the TT. We found a similar misclassification in the TT class, where 37 and 42 samples predicted the TV and TC, respectively. In addition, the TC class images feature 29 images overlapped with TT. The model wrongly predicted about $6-8\%$ samples to TC and TT categories. We also measured the class-wise AUC scores shown in Fig. 13. All four brain anatomy classes achieved higher AUC scores greater than 97%.

To classify brain anatomy, the last layer embedding of the *COMFormer* is a 4-dimensional class vector that maps into 2 dimensions for visualizing the t-SNE in Fig. 14. The resulting plot reveals that the NB and TC classes are distinct and form separate clusters. However, the feature space for the TT and TV classes overlaps significantly.

Fig. 15 illustrates three instances of misclassification, where the classes NB, TC, and TV were incorrectly predicted as TT. The NB example contains some outer spatial brain structures



(a) (b) (c)

Fig. 15: Samples misclassified as TT by the *COMFormer* model. Here (a), (b), and (c) belong to the ground-truth (GT) labels of NB, TC, and TV brain anatomy classes, respectively.

with hyperechoic circular boundaries that are similar to the TT class features. Similarly, the TC class example appears to TT, and the model misclassifies it. In contrast, the TV image has poor contrast and displays acoustic shadows, which the model incorrectly identified as TT.

D. Discussion and limitations

We have proposed a transformer-based *COMFormer* architecture to classify the maternal-fetal and brain anatomical structures from US images. For this purpose, we designed a residual cross-covariance attention block that computes the attention in channels or feature dimensions. We found that previously studied transformers used their self-attention mechanism which does not aggregate the local feature information and required significantly more training samples to achieve better classification performance than CNNs. We utilized a publicly available fetal US dataset, in which both different sub-classes samples are very distinct based on their anatomical structures. A common architecture was in building two models. The first model was applied to maternal-fetal anatomical structures classification that achieved higher results by extracting local and global features where other compared methods failed to deal with challenging artefacts. The heat maps of the proposed model focused on targeted anatomical structures while ignoring the artefacts such as acoustic shadows and neighbouring hyperechoic tissues. We also assess the strength of *COMFormer* architecture for the brain anatomy classification task. Our experimental results surpassed the recent work [18] by 2.5% for classifying the six class maternal-fetal anatomical structures. However, we also compared our model results for brain anatomy with [18] and achieved a higher accuracy of 11%.

The study's limitations arise from the lack of heterogeneity in the fetal ultrasound dataset, which introduces potential bias and limits the generalizability of the findings. The resulting classification model may not perform well across diverse populations, resulting in reduced accuracy and reliability in clinical settings. Additionally, the single dataset may fail to capture the full range of fetal structures, anatomical variations, and developmental anomalies, thereby limiting the applicability of the study's conclusions. To improve the accuracy, reliability, and inclusiveness of fetal ultrasound interpretations, it is crucial to obtain a more diverse and representative dataset. Furthermore, it has come to our attention that this study needs more comprehensive coverage of crucial fetal anatomical aspects since we only considered limited maternal fetal and brain anatomy structures without the inclusion of a specific 'None' class that represents the absence of any kind of anatomical structure in an image. As a result, our future endeavors will be dedicated to addressing these significant gaps.

IV. Conclusion

This paper presents the transformer-based *COMFormer* architecture to classify maternal-fetal and brain anatomical structures in 2D US images. We designed a new residual cross-covariance attention (R-XCA) block that improved the classification results from the original cross-covariance attention layers. The *COMFormer* model has been trained and validated on a publicly available dataset called "BCNatal" with models built for two classification tasks. We have presented a broad range of justification experiments that present the effectiveness of the proposed model. By extracting complex features from challenging US images, the *COMFormer* model has exceeded the limitations of several CNN and transformer-based models, resulting in highly accurate results. The suggested model accurately identifies the maternal-fetal and brain structures that could help monitor the healthy development of fetuses. Future work will explore the applicability of the proposed approach for fetal US image classification for different trimesters and on-scan video clips to incorporate the temporal aspect of real-time ultrasound scans.

Acknowledgments

This work was supported in part by the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography).

References

[1] S. Płotka, A. Klasa, A. Lisowska, J. Seliga-Siwecka, M. Lipa, T. Trzciński, and A. Sitek, "Deep learning fetal ultrasound video model match human observers in biometric measurements," *Physics in Medicine & Biology*, vol. 67, no. 4, p. 045013, 2022.

[2] M. A. Maraci, R. Napolitano, A. Papageorghiou, and J. A. Noble, "Object classification in an ultrasound video using lp-sift features," in *International MICCAI Workshop on Medical Computer Vision.* Springer, 2014, pp. 71–81.

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

[4] S. Gofer, O. Haik, R. Bardin, Y. Gilboa, and S. Perlman, "Machine learning algorithms for classification of first-trimester fetal brain ultrasound images," *Journal of Ultrasound in Medicine*, vol. 41, no. 7, pp. 1773–1779, 2022.

[5] C. P. Bridge, C. Ioannou, and J. A. Noble, "Automated annotation and quantitative description of ultrasound videos of the fetal heart," *Medical image analysis*, vol. 36, pp. 147–161, 2017.

[6] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE journal of biomedical and health informatics*, vol. 19, no. 5, pp. 1627–1636, 2015.

[7] M. Yaqub, B. Kelly, A. T. Papageorghiou, and J. A. Noble, "Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2015, pp. 687–694.

[8] 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 transactions on medical imaging*, vol. 27, no. 9, pp. 1342–1355, 2008.

[9] M. Yaqub, N. Sleep, S. Syme, Z. Chen, H. Ryou, S. Walton, J. A. Noble, and A. T. Papageorghiou, “491 scannav® audit: an ai-powered screening assistant for fetal anatomical ultrasound,” *American Journal of Obstetrics & Gynecology*, vol. 224, no. 2, p. S312, 2021.

[10] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, “Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound,” *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[11] N. Savioli, E. Grisan, S. Visentin, E. Cosmi, G. Montana, and P. Lamata, “Real-time diameter of the fetal aorta from ultrasound,” *Neural Computing & Applications*, vol. 32, pp. 6735 – 6744, 2019.

[12] R. Yasrab, Z. Fu, H. Zhao, L. H. Lee, H. Sharma, L. Drukker, A. T. Papageorgiou, and J. A. Noble, “A machine learning method for automated description and workflow analysis of first trimester ultrasound scans.” *IEEE Transactions on Medical Imaging*, 2022.

[13] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng, “Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 507–514.

[14] L. H. Lee, Y. Gao, and J. A. Noble, “Principled ultrasound data augmentation for classification of standard planes,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 729–741.

[15] M. Ahmed and J. A. Noble, “An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1084–1087.

[16] Y. Cai, H. Sharma, P. Chatelain, and J. A. Noble, “Sonoeyenet: standardized fetal ultrasound plane detection informed by eye tracking,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1475–1478.

[17] J. Xi, J. Chen, Z. Wang, D. Ta, B. Lu, X. Deng, X. Li, and Q. Huang, “Simultaneous segmentation of fetal hearts and lungs for medical ultrasound images via an efficient multi-scale model integrated with attention mechanism,” *Ultrasonic Imaging*, vol. 43, no. 6, pp. 308–319, 2021.

[18] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós, “Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes,” *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[19] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, “Transformers in medical image analysis: A review,” *Intelligent Medicine*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667102622000717

[20] B. Gheflati and H. Rivaz, “Vision transformers for classification of breast ultrasound images,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 480–483.

[21] S. Płotka, M. K. Grzeszczyk, R. Brawura-Biskupski-Samaha, P. Gutaj, M. Lipa, T. Trzciński, and A. Sitek, “Babynet: Residual transformer module for birth weight prediction on fetal ultrasound video,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 350–359.

[22] C. Yang, S. Liao, Z. Yang, G. Jiaqi, Z. Zhang, Y. Yingjian, Y. Guo, S. Yin, C. Liu, and Y. Kang, “Rdhcformer: Fusing resdcn and transformers for fetal head circumference automatic measurement in 2d ultrasound images,” *Frontiers in Medicine*, p. 839, 2022.

[23] I. Sarris, C. Ioannou, P. Chamberlain, E. Ohuma, F. Roseman, L. Hoch, D. G. Altman, A. T. Papageorghiou, and for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) , “Intra- and interobserver variability in fetal ultrasound measurements,” *Ultrasound in Obstetrics & Gynecology*, vol. 39, no. 3, pp. 266–273, 2012. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.10082

[24] N. J. Dudley and E. Chapman, “The importance of quality management in fetal measurement,” *Ultrasound in Obstetrics & Gynecology*, vol. 19, no. 2, pp. 190–196, 2002. [Online]. Available: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1046/j.0960-7692.2001.00549.x

[25] M. Ahmed and J. A. Noble, “Fetal ultrasound image classification using a bag-of-words model trained on sonographers’ eye movements,” *Procedia Computer Science*, vol. 90, pp. 157–162, 2016.

[26] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, “A review on deep-learning algorithms for fetal ultrasound-image analysis,” *Medical Image Analysis*, p. 102629, 2022.

[27] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, “Xcit: Cross-covariance image transformers,” *Advances in neural information processing systems*, vol. 34, 2021.

[28] M. A. Hassanien, V. K. Singh, D. Puig, and M. Abdel-Nasser, “Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences,” *Diagnostics*, vol. 12, no. 5, p. 1053, 2022.

[29] V. K. Singh, B. Kucukgoz, D. C. Murphy, X. Xiong, D. H. Steel, and B. Obara, “Benchmarking automated detection of the retinal external limiting membrane in a 3d spectral domain optical coherence tomography image dataset of full thickness macular holes,” *Computers in Biology and Medicine*, vol. 140, p. 105070, 2022.

[30] S. Bezryadin, P. Bourov, and D. Ilinih, “Brightness calculation in digital image processing,” in *International Symposium on Technologies for Digital Photo Fulfillment*, vol. 2007, no. 1, 2007, pp. 10–15.

[31] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[33] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[34] 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.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[36] 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, pp. 4700–4708.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[38] H. Bao, L. Dong, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.

[39] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, “Going deeper with image transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 32–42.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[42] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.