Anatomy-Aware Contrastive Representation Learning for Fetal Ultrasound

Zeyu Fu[1⋆], Jianbo Jiao[1⋆], Robail Yasrab[1⋆], Lior Drukker[2,3], Aris T. Papageorghiou[2], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
zeyu.fu@eng.ox.ac.uk
[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK
[3] Department of Obstetrics and Gynecology, Tel-Aviv University, Israel

Abstract. Self-supervised contrastive representation learning offers the advantage of learning meaningful visual representations from unlabeled medical datasets for transfer learning. However, applying current contrastive learning approaches to medical data without considering its domain-specific anatomical characteristics may lead to visual representations that are inconsistent in appearance and semantics. In this paper, we propose to improve visual representations of medical images via anatomy-aware contrastive learning (AWCL), which incorporates anatomy information to augment the positive/negative pair sampling in a contrastive learning manner. The proposed approach is demonstrated for automated fetal ultrasound imaging tasks, enabling the positive pairs from the same or different ultrasound scans that are anatomically similar to be pulled together and thus improving the representation learning. We empirically investigate the effect of inclusion of anatomy information with coarse- and fine-grained granularity, for contrastive learning and find that learning with fine-grained anatomy information which preserves intra-class difference is more effective than its counterpart. We also analyze the impact of anatomy ratio on our AWCL framework and find that using more distinct but anatomically similar samples to compose positive pairs results in better quality representations. Extensive experiments on a large-scale fetal ultrasound dataset demonstrate that our approach is effective for learning representations that transfer well to three clinical downstream tasks, and achieves superior performance compared to ImageNet supervised and the current state-of-the-art contrastive learning methods. In particular, AWCL outperforms ImageNet supervised method by 13.8% and state-of-the-art contrastive-based method by 7.1% on a cross-domain segmentation task. The code is available at https://github.com/JianboJiao/AWCL.

Keywords: Representation learning · Contrastive learning · Ultrasound.

⋆ Equal contribution.

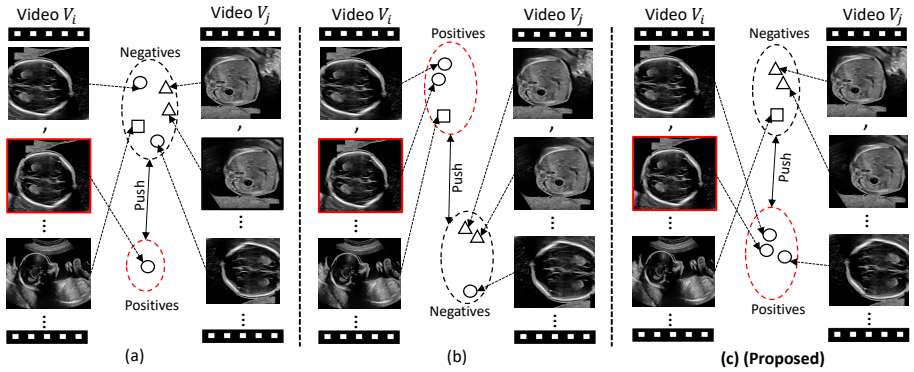

Fig. 1. Illustration of different representation learning approaches for fetal ultrasound, including (a) self-supervised contrastive learning, (b) contrastive learning with patient metadata, and (c) our proposed anatomy-aware contrastive learning. Icon shapes of circle (○), square (□) and triangle (△) denote the anatomical categories of fetal head, profile, and abdomen, respectively. The anchor image is highlighted with a red bounding box, while the red dotted circle means pull together (Best viewed in colored version).

1 Introduction

Semi-supervised and self-supervised representation learning with or without annotations have attracted significant attention across various medical imaging modalities [7, 10, 26–28]. These learning schemes are able to well exploit large-scale unlabeled medical datasets and learn meaningful representations for downstream task finetuning. In particular, contrastive representation learning based on instance discrimination tasks [6, 11] has become the leading paradigm for self-supervised pretraining, where a model is trained to pull together each instance and its augmented views and meanwhile push it away from those of all other instances, in the embedding space.

However, directly applying self-supervised contrastive learning (e.g. SimCLR [6] and MoCo [11]) in the context of medical imaging may result in visual representations that are inconsistent between appearances and semantics. We illustrate this issue in Fig. 1(a), which shows that vanilla contrastive learning approach without considering the domain-specific anatomical characteristics lead to false negatives, i.e. some negative samples having high affinity with the anchor image are pushed away. To address this, we explore the following question: *Is domain-specific anatomy information helpful in learning better representations for medical data?*

We investigate this question via the proposed anatomy-aware contrastive learning (AWCL), as depicted in Fig. 1(c), where "anatomy-aware" here means that the inclusion of anatomy information is leveraged to augment the positive/negative pair sampling in a contrastive learning manner. In this work, we demonstrate the proposed approach for fetal ultrasound imaging tasks, where a number of different fetal anatomies can be present in a diagnostic scan. Moti-

vated by Khosla et al. [18], we expand the pool of positive samples by grouping images from the same or different ultrasound scans that share common anatomical categories. More importantly, our approach is optimized alternately with both conventional and anatomy-aware contrastive objectives, as shown in Fig. 2(a), given that the anatomy information is not always accessible for each sampling process. Moreover, we consider both coarse- and fine-grained anatomical categories with the availability for data sampling, as shown in Fig. 2(b) and (c). We also empirically investigate their effect on the transferability of the learned feature representations. To assess the effectiveness of our pre-trained representations, we perform evaluation of transfer learning on three downstream clinical tasks: standard plane detection, segmentation of Crown Rump Length (CRL) and Nuchal Translucency (NT), and recognition of first-trimester anatomical structures. In summary, the main contributions and findings are:

- We develop an anatomy-aware contrastive learning approach for medical fetal ultrasound imaging tasks.
- We empirically compare the effect of inclusion of anatomy information with coarse- and fine-grained granularity respectively, within our contrastive learning approach. The comparative analysis suggests that contrastive learning with fine-grained anatomy information which preserves intra-class difference is more effective than its counterpart.
- Extensive experimental evaluations on three downstream clinical tasks demonstrate the better generalizability of our proposed approaches over learning from supervised ImageNet data, vanilla contrastive learning [6], and contrastive learning with patient information [2, 7, 25].
- We provide an in-depth analysis to show the proposed approach learns high-quality discriminative representations.

2 Related work

Self-supervised learning (SSL) in medical imaging. Prior works using SSL for medical imaging typically rely on designing pre-text tasks, such as addressing a Rubik' cube [28], image restoration [14, 27], predicting anatomical position [3] and multi-task joint reasoning [17]. Recently, contrastive based SSL [6, 11] has been favourably applied to learn more discriminative representations across various medical imaging tasks [7, 24, 26]. In particular, Sowrirajan et al. [24] successfully adapted a MoCo-contrastive learning method [11] into chest X-rays and demonstrated better transferable representations and initialization for chest X-ray diagnostic tasks. Taher et al. [13] presented a benchmark evaluation study to investigate the effectiveness of several established contrastive learning models pre-trained on ImageNet on a variety of medical imaging tasks. In addition, there have been recent approaches [2, 7, 25] that leverage patient metadata to improve the medical imaging contrastive learning. These approaches constrained the selection of positive pairs only from the same subject (video), with the assumption that visual representations from the same subject share similar semantic meaning. However, these approaches may not generalize well to a scenario,

where different organs or anatomical structures are captured in a single video. For instance, as seen from Fig. 1(b), some positive pairs having low affinity in visual appearance and semantics are pulled together, i.e. false positives, which degrades the representation learning. The proposed learning scheme, as shown in Fig. 1(c), is advantageous to address the aforementioned limitations by augmenting the sampling process with the inclusion of anatomy information. Moreover, our approach differs from [26] and [7] which combine label information as an additional supervision signal with self supervision for multi-tasking.

Representation learning in fetal ultrasound. There are related works exploring representation learning for fetal ultrasound imaging tasks. Baumgartner et al. [4] and Schlemper et al. [22] proposed a VGG-based network and an attention-gated network respectively to detect fetal standard planes. Sharma et al. [23] presented a multi-stream network which combines 2D image and spatio-temporal information to automate clinical workflow description of full-length routine fetal anomaly ultrasound scans. Cai et al. [5] considered incorporating the temporal dimension into visual attention modelling via multi-task learning for standard biometry plane-finding navigation. However, the generalization and transferability of those models to other target tasks remains unclear. Droste et al. [8] proposed to learn transferable representations for fetal ultrasound interpretation by modelling sonographer visual attention (gaze tracking) without manual supervision. More recently, Jiao et al. [16] proposed to derive a meaningful representation from raw data by developing a cross-modal contrastive learning which aligns the correspondence between fetal ultrasound video and narrative speech audio. Our work differs by focusing on learning general image representations without requiring additional data modalities (e.g. gaze tracking and audio) from the domain of interest, and we also perform extensive experimental analysis on three downstream clinical tasks to assess the effectiveness of the learned representations.

3 Fetal Ultrasound Imaging Dataset

This study uses a large-scale fetal ultrasound imaging dataset , which was acquired as part of PULSE (Perception Ultrasound by Learning Sonographic Experience) project [9]. This dataset considered for pre-training contains full-length routine second-trimester (gestational age of 18–22 weeks) fetal ultrasound videos, recorded from the ultrasound scanner display using a lossless compression and sampled at the rate of 30 Hz. The scans were performed by operators including sonographers and fetal medicine specialists using a commercial Voluson E8 version BT18 (General Electric, Zipf, Austria) ultrasound machine. During a routine scanning session, the operator views several fetal or maternal anatomical structures. The frozen views saved by sonographers are referred to as *standard planes* in the paper, following the UK Fetal Anomaly Screening Programme (FASP) nomenclature [1].

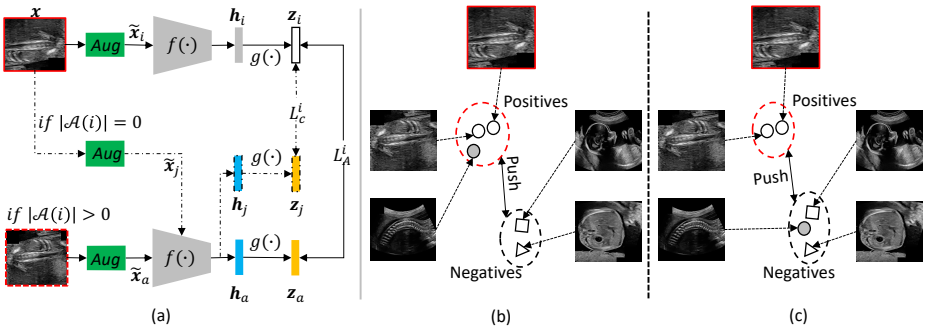

Fig. 2. (a) presents the overview of proposed anatomy-aware contrastive learning approach. (b) and (c) illustrate using coarse and fine-grained anatomy categories, respectively for the proposed AWCL framework. Icon shapes of white-circle (○), grey-circle (⊙), square (□) and triangle (△) denote the classes of coronal view of spine, sagittal view of spine, profile, and abdomen, respectively.

We consider a subset of the entire ultrasound dataset for the proposed pre-training approach, which consists of total number of 2,050,432 frames[4] from 534 ultrasound videos. In this sub-dataset, there are 15,384 frames labeled with 13 fine-grained anatomy categories, including four views of heart, three-vessel and trachea (3VT), four-chamber (4CH), right ventricular outflow tract (RVOT), and left ventricular outflow tract (LVOT), two views of brain, transventricular (BrainTv.) and transcerebellum (BrainTc.), two views of spine, coronal (SpineCor.) and sagittal (SpineSag.), abdomen, femur, kidneys, lips, profile and background class. In addition, 69,671 frames are labeled with coarse anatomy categories without dividing the heart, brain and spine into further sub-categories as those of above, but also 3D mode, maternal anatomy including Doppler, abdomen, nose and lips, kidneys, face-side profile, full-body-side profile, bladder including Doppler, femur and "Other" class. All Image frames were preprocessed by cropping the ultrasound image region and resizing it to 224×288 pixels.

4 Method

In this section, we first describe the problem formulation of contrastive learning with medical images, and then present our anatomy-aware contrastive learning algorithm design as well as training details.

4.1 Problem formulation

For each input image $\mathbf{x}$ in a mini-batch of N samples, randomly sampled from a pre-training dataset $\mathcal{V}$, a contrastive learning framework (i.e. SimCLR [6]) applies two augmentations to obtain a positive pair $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$, yielding a set of $2N$

[4] Every 8th frame is extracted to reduce temporal redundancy of ultrasound videos.

samples. Let i denote the anchor input index, the contrastive learning objective can be defined as,

$$L_C^i = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}, \tag{1}$$

where $\mathbf{1} \in \{0, 1\}$, τ is a temperature parameter and $sim(\cdot)$ is the pairwise cosine similarity. $\mathbf{z}$ is a representation vector, calculated by $\mathbf{z} = g(f(\mathbf{x}))$, where $f(\cdot)$ denotes a shared encoder modelled by a convolutional neural network (CNN) and $g(\cdot)$ is a multi-layer perception (MLP) projection head.

The above underpins the vanilla contrastive learning of medical images. However in some cases (e.g. ultrasound scan as illustrated in this paper), this standard approach, as well as its extended version by leveraging the patient information [2,7,25], may lead to false negatives and false positives respectively, as seen from Fig. 1(a) and (b). To this end, we introduce a new approach as detailed below.

4.2 Anatomy-aware contrastive learning

Fig. 1(c) illustrates the main idea of the new anatomy-aware contrastive learning (AWCL) approach, which incorporates additional samples from same or different US scans belonging to the same anatomy category. In addition to positive sampling from the same image and its augmentation, AWCL is tailored to the case where multiple anatomical structures are present.

As shown in Fig. 2(a), we utilize the available anatomy information as detailed in Section 3, forming a positive sample set $\mathcal{A}(i)$ with the same anatomy as sample i. The assumption for such a design is that image samples within the same anatomy category should have similar appearances, based on a clinical perspective [9]. Motivated by [18], we design the anatomy-aware contrastive learning objective as follows,

$$L_A^i = -\frac{1}{|\mathcal{A}(i)|} \sum_{a\in\mathcal{A}(i)} \log \frac{\exp\left(\text{sim}\left(z_i, z_a\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\text{sim}\left(z_i, z_k\right)/\tau\right)}, \tag{2}$$

where $|\mathcal{A}(i)|$ denotes the cardinality.

Due to the limited availability of anatomical categories, $\mathcal{A}(i)$ is not always achievable for each sampling process. In this regard, the AWCL framework is formulated as an alternate optimization combining both learning objectives of Eq. 1 and Eq. 2. This gives a loss function defined as

$$L^i = \begin{cases} L_C^i & \text{if } |\mathcal{A}(i)| = 0 \\ L_A^i & \text{if } |\mathcal{A}(i)| > 0. \end{cases} \tag{3}$$

Furthermore, we consider both coarse- and fine-grained anatomical categories for the proposed AWCL framework, and compare their effect on the transferability of visual representations. Fig. 2(b) and (c) shows the motivation of this

Algorithm 1: Anatomy-aware Contrastive Learning (AWCL)

Input : Sample x_i and its positive set $\mathcal{A}(i)$, pre-training dataset $\mathcal{V}$
Output: The loss value L of the current learning step

Sample mini-batch training data $x_i \in \mathcal{V}$
if $|\mathcal{A}(i)| = 0$ **then**
  Apply data augmentations $\rightarrow$ positive pair $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$
  Extract representation vectors $z_i = g(f(\tilde{\mathbf{x}}_i)), z_j = g(f(\tilde{\mathbf{x}}_j))$
  $L = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$
else
  Sample data x_a with the same anatomy as x_i, where $x_a \in \mathcal{A}(i)$
  Apply data augmentations $\rightarrow$ positive pair $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_a)$
  Extract representation vectors $z_i = g(f(\tilde{\mathbf{x}}_i)), z_a = g(f(\tilde{\mathbf{x}}_a))$
  $L = -\frac{1}{|\mathcal{A}(i)|} \sum_{a \in \mathcal{A}(i)} \log \frac{\exp(\text{sim}(z_i, z_a)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$
end
Return L

comparative analysis. For an anatomical structure with different views of visual appearance (e.g. the spine has two views as sub-classes), we observe that AWCL with coarse-grained anatomy information tends to minimize the intra-class difference by pulling together all the instances of the same anatomy. In contrast, AWCL with fine-grained anatomy information tends to preserve the intra-class difference by pushing away images with different visual appearances despite the same anatomy. Both strategies of the proposed learning approach are evaluated and compared in Section 6.3. We further study the impact of the ratio of anatomy information used in AWCL pre-training in Section 6.4.

4.3 Implementation details

Algorithm 1 provides the pseudo-code of AWCL. Following the prior arts [7, 24, 25], we use ResNet-18 [12] as our backbone architecture. Further studies on different network architectures are out of scope of this paper. We split the pre-training dataset as detailed in Section 3 into training and validation sets (80%/20%), and train the model using the Adam optimizer with a weight decay of 10^{-6}, and a mini-batch size of 32. We follow [6] for the data augmentations applied to the sampled training data. The output feature dimension of z is set to 128. The temperature parameter τ is set as 0.5. The models are trained with the loss functions defined earlier (Eq. 2 and Eq. 1) for 10 epochs. The learning rate is set as 10^{-3}. The whole framework is implemented with the PyTorch [21] framework on a PC with NVIDIA Titan V GPU card.

To demonstrate the effectiveness of our AWCL trained models, we compare them with random initialization, ImageNet pre-trained ResNet18 [12], supervised pre-training with coarse labels, supervised pre-training with fine-grained labels, vanilla contrastive learning (SimCLR) [6], and contrastive learning with patient

Table 1. Details of the downstream datasets and imaging tasks.

Trimester	Task	#Scans	#Images	#Classes
2nd	I- Standard Plane Detection	58	1,075	14
1st	II- Recognition of first-trimester anatomies	90	25,490	5
1st	III- Segmentation of NT and CRL	128	16,093	3

information (CLPI) [2, 7, 19]. All pre-training methods present here are pre-trained from scratch on the pre-training dataset with the similar parameter configurations as listed above .

5 Experiments on Transfer Learning

In this section, we evaluate pre-trained representations by supervised transfer learning with end-to-end fine-tuning on three downstream clinical tasks, which are standard plane detection (Task I), recognition of first-trimester anatomies (Task II) and segmentation of NT and CRL (Task III). The datasets for downstream task evaluation are listed in Table 1, and come from a large-scale fetal ultrasound dataset [9]. For fair comparison, all compared pre-training methods were fine-tuned with the same parameter settings and data augmentation policies within each downstream task evaluation.

5.1 Evaluation on standard plane detection

Evaluation details. Here, we investigate how the pre-trained representations generalize to a in-domain second-trimester classification task, which consists of the same fine-grained anatomical categories as detailed in Section 3. We fine-tune each trained backbone encoder and attach a classifier head [4] to train the entire network for 70 epochs with a learning rate of 0.01, decayed by 0.1 at epochs 30 and 55. The network training is performed via SGD with momentum of 0.9, weight decay of 5×10^{-4}, mini-batch size of 16 and a cross-entropy loss, and it is evaluated via a three-fold cross validation. The augmentation policy used is analogous to [8], including random horizontal flipping, rotation (10 degrees), and varying gamma and brightness. We employ precision, recall and F1-scores computed as macro-averages as the evaluation metrics.

Results and discussion. Table 2 shows a quantitative comparison of fine-tuning performance on the three evaluated downstream tasks. From the results of Task I, we observe that our pre-trained models, i.e. *AWCL (coarse)* and *AWCL (fine-grained)*, generally outperform the compared contrastive learning methods SimCLR and CLPI. In particular, our *AWCL (coarse)* improves SimCLR and CLPI by 1.9% and 3.8% in F1-score, respectively. Compared to the supervised pre-training methods, both proposed approaches achieve better performance in Recall and F1-score than vanilla supervised pre-training with coarse-grained labels. These findings validate the efficacy of our proposed approaches and suggest

Table 2. Quantitative comparison of fine-tuning performance (mean±std.[%]) on the tasks of standard plane detection (Task I), first-trimester anatomy recognition (Task II) and CRL / NT segmentation (Task III). Best results are marked in **bold.**

	Task I			Task II			Task III		
Pre-training methods	Precision (↑)	Recall (↑)	F1-score (↑)	Precision (↑)	Recall (↑)	F1-score (↑)	GAA (↑)	MA(↑)	mIoU(↑)
Rand.Init.	70.4±1.7	58.3±3.1	61.6±3.1	81.4±3.4	79.2±0.1	81.5±0.3	67.3±0.2	63.0±2.1	46.7±0.1
ImageNet	78.8±4.6	73.6±4.1	73.6±2.8	92.0±0.5	93.4±1.5	92.1±2.9	71.6±1.3	64.2 ± 1.0	49.0±0.1
Supervised (coarse)	74.2±2.7	67.5±3.4	69.0±3.2	95.2±0.1	93.7±0.2	94.1±0.4	76.4±0.3	67.5±1.1	50.1±0.3
Supervised (fine-grained)	**84.6±1.0**	**77.1±2.3**	**78.6±2.1**	96.1±0.1	96.8±1.0	96.4±0.9	80.0±0.2	75.5±0.1	62.8±0.4
SimCLR	71.7±0.3	69.6±1.5	69.4±0.7	96.0±0.5	95.2±0.3	95.2±0.8	77.6±1.4	69.2±0.1	55.7±0.2
CLPI	68.6±4.2	68.5±3.2	67.5±3.7	89.2±0.1	88.3±0.8	89.6±1.1	72.7±0.2	65.4±1.4	48.1±1.2
AWCL (coarse)	71.4±3.3	73.1±1.9	71.3±2.2	95.6±0.7	96.2±1.6	95.9±0.2	79.8±0.7	76.1±0.3	61.2±1.3
AWCL (fine-grained)	71.8±2.7	70.0±1.2	70.1±1.7	**96.9±0.1**	**96.8±1.8**	**97.1±0.2**	**80.2±1.1**	**76.0±0.5**	**62.8±0.1**

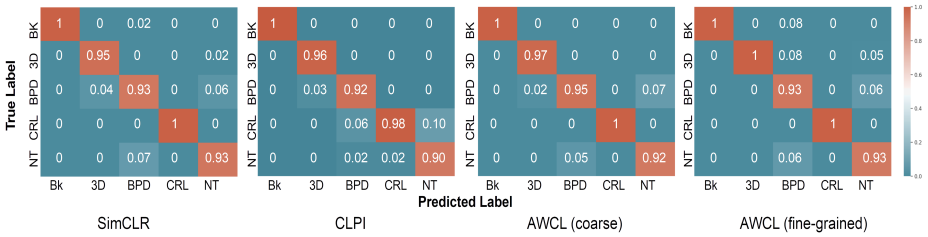

Fig. 3. Illustration of the confusion matrix for the first-trimester classification task.

that incorporating anatomy information to select positive pairs from multiple scans can notably improve representation learning.

However, we find that all the contrastive pre-training approaches present here underperform the supervised pre-training (fine-grained) which has the same form of semantic supervision as Task I. This may suggest that without explicitly encoding semantic information, contrastively learned representations may provide limited benefits to the generalization of a fine-grained multi-class classification task, which is line with the findings in [15].

5.2 Evaluation on recognition of first-trimester anatomies

Evaluation details. We investigate how the pre-trained representations generalize to a cross-domain classification task using the first-trimester US scans. This first-trimester classification task recognises five anatomical categories: crown rump length(CRL), nuchal translucency (NT), biparietal diameter (BPD), 3D and background (Bk). We split the data into training and testing sets (78%/22%). The trained encoders followed by two fully connected layers and a softmax layer were fine-tuned for 200 epochs with a learning rate of 0.1 decayed by 0.1 at 150 epochs. The network was trained using SGD with momentum of 0.9. Standard data augmentation was used, including rotation $[-30°, 30°]$, horizontal flip, Gaussian noise, and shear ≤ 0.2. Batch size was adjusted according to model size and GPU memory restrictions. We use the same metrics as presented in Task I for performance evaluation.

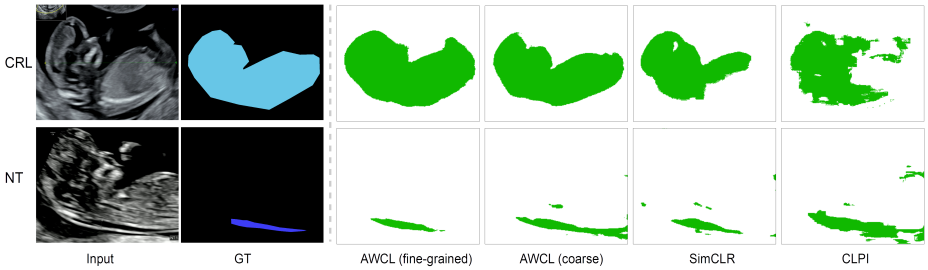

Fig. 4. Illustration of the qualitative results for the first-trimester segmentation task.

Results and discussion. For Task II, we can see from Table 2, our approach *AWCL (fine-grained)* achieves the best performance among all the compared solutions. In particular, it achieves a performance gain of 4.9%, 3.4% and 5.0% in Precision, Recall and F1-score compared to ImageNet, and even improves supervised pre-training with fine-grained labels (the upper-bound baseline) by 0.7% in F1-score. Moreover, our proposed *AWCL (coarse)* also surpasses ImageNet and supervised pre-training with coarse-grained labels by 1.9% and 6.3% in F1-score. For comparison with other contrastive learning methods, we observe a similar improved trend as described in Task I, i.e. our methods *AWCL (coarse)* and *AWCL (fine-grained)* perform better than SimCLR and CLPI. This can be further validated from Fig. 3, which shows that both *AWCL (coarse)* and *AWCL (fine-grained)* provide better prediction accuracy than CLPI in all anatomy categories. These experimental results again demonstrate the effectiveness of our AWCL approaches and suggest that the inclusion of anatomy information in contrastive learning is a good practice when it is available at hand.

5.3 Evaluation on segmentation of NT and CRL

Evaluations details. In this section, we evaluate how the pre-trained models generalize to a cross-domain segmentation task with the data from the first-trimester US scans. Segmentation of NT and CRL was defined as a three-class segmentation task with the three classes being; mid-sagittal view, nuchal translucency, background. The data is divided into training and testing with 80%/20%. We follow the design of ResNet-18 auto-encoder by attaching additional decoders with the trained encoders, and then fine-tuned the entire model for 50k iterations with a learning rate of 0.001, RMSprop optimization (momentum=0.9) and a weight decay of 0.001. We apply random scaling, random shifting, random rotation, and random horizontal flipping for data augmentation. We use global average accuracy (GAA), mean accuracy (MA), and mean intersection over union (mIoU) metrics for evaluating the segmentation task (Task III).

Results and discussion. For Task III, we find that our proposed *AWCL (fine-grained)*, achieves comparable or slightly better performance than supervised pre-training with fine-grained labels and surpasses other compared pre-training

methods by large margins in mIoU (see Table 2). In particular, it outperforms ImageNet and SimCLR by 13.8% and 7.1% in mIoU, respectively. Likewise, our proposed *AWCL (coarse)* performs better than ImageNet, supervised pre-training with coarse-grained labels, SimCLR and CLPI by large margins in most evaluation metrics. Fig. 4 also visualizes the superior performance of *AWCL (fine-grained)* and *AWCL (coarse)* compared to SimCLR and CLPI, which aligns with the quantitative evaluation. These observations suggest that our approaches are able to learn more meaningful semantic representations that are beneficial for pixel-wise segmentation task. Overall, the evaluated results on Tasks II and III demonstrate that our AWCL approaches report consistently better performance than the compared pre-training approaches, implying the advantage of learning task-agnostic features that are better generalized to the tasks from different domains.

6 Analysis

6.1 Partial fine-tuning

To analyze the representation quality, we extract fixed feature representations from the last layer of ResNet-18 encoder and then evaluate them in two classification target tasks (Task I and Task II). Experimentally, we freeze the entire backbone encoder and attach a classification head [4] for Task I, and a non-linear classifier as mentioned in Section 5.2 for Task II. From Table 3, we observe that our AWCL approaches have shown better representation quality by surpassing three compared approaches in F1-score in both tasks. This suggests that the our learned representations are strong non-linear features which are more generalizable and transferable to the downstream tasks. By comparing the results between Tables 2 and 3, we find that although the reported scores of partial fine-tuning are generally lower than those as in full fine-tuning, the performance between two implementations of transfer learning is correlated.

Table 3. Performance comparison of partial fine-tuning (mean±std.[%]) on the tasks of standard plane detection (Task I) and first-trimester anatomy recognition (Task II). Best results are marked in **bold.**

	Task I			Task II		
Pre-training methods	Precision (↑)	Recall (↑)	F1-score (↑)	Precision (↑)	Recall (↑)	F1-score (↑)
ImageNet	65.5±4.9	58.1±4.3	60.2±4.9	84.03±0.13	84.25±0.45	83.92±0.62
SimCLR	67.6±3.5	67.3±4.1	65.9±3.6	87.65±0.09	86.82±0.11	86.12±0.20
CLPI	**71.2±0.6**	68.6±4.9	67.9±5.1	82.07±0.62	80.28±0.83	81.10±1.03
AWCL (coarse)	70.5±2.7	**71.3±1.7**	**69.5±1.9**	86.21±1.20	87.67±0.32	86.14±0.59
AWCL (fine-grained)	69.7±1.0	68.8±0.2	68.4±0.5	**88.65±0.49**	**88.14±0.17**	**87.00±0.01**

6.2 Visualization of feature representations

In this section, we investigate why the feature representations produced from our approaches result in better downstream task performance. We visualize the

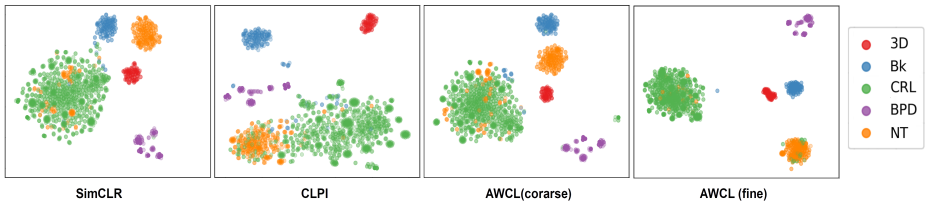

Fig. 5. t-SNE feature visualization of the model penultimate layers on Task II.

image representations of Task II extracted from the penultimate layers using t-SNE [20] in Fig. 5, where different anatomical categories are denoted with different color. We compare the resulting t-SNE embeddings of our approaches with those as SimCLR and CLPI. We observe that the feature representation by CLPI is not quite separable, especially for classes of NT (orange) and CRL (green). The features embeddings from SimCLR are generally better separated than those in CLPI, while confusion between CRL (green) and Bk (blue) remains. By comparison, *AWCL (fine-grained)* achieves the best separated clusters among five anatomical categories, which means that the learned representations in the embedding space are more distinguishable. These visualization results demonstrate that our proposed approaches are able to learn discriminative feature representations which are better generalized to downstream tasks.

6.3 Impact of data granularity on AWCL

We analyze how the inclusion of coarse- and fine-grained anatomy information impact on our AWCL framework, by comparing the experimental results between *AWCL (coarse)* and *AWCL (fine-grained)* from Section 5.1 to Section 6.2. Based on the transfer learning results in Table 2, we find that *AWCL (fine-grained)* achieves better performance than *AWCL (coarse)* in Tasks II and III, despite the slight performance drop in Task I. We hypothesize that *AWCL (coarse)* learns more generic representations than *AWCL (fine-grained)*, which leads to better in-domain generalization performance. Qualitative results in Fig. 3 and Fig. 4 also reveal the advantage of *AWCL (fine-grained)* over its counterpart. Based on the ablation analysis, Table 3 shows a similar finding as observed in Table 2. Fig. 6 shows that feature embeddings of *AWCL (fine-grained)* are more discriminative than those of *AWCL (coarse)* thereby resulting in better generalization performance to downstream tasks. These observations suggest the importance of learning more intra-class feature representations for better generalization to downstream tasks especially when there is domain shift.

6.4 Impact of anatomy ratio on AWCL

We investigate how varying anatomy ratios impact on our AWCL framework. Note that higher anatomy ratio represents that larger number of samples from same or different US scans belonging to the same anatomy category are included

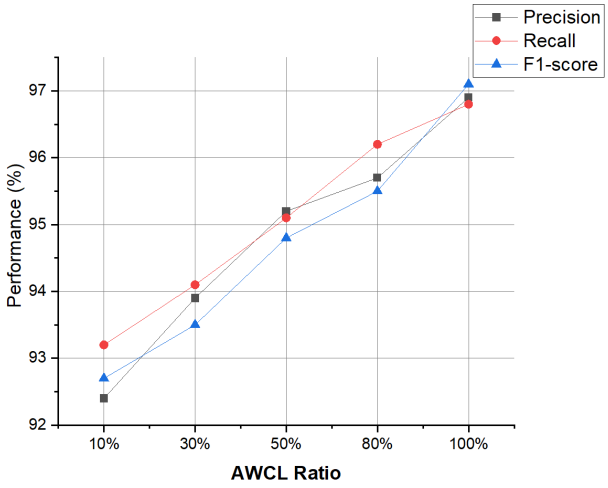

Fig. 6. Impact of anatomy ratio on AWCL (fine-grained) evaluated on Task II.

to form positive pairs for contrastive learning. In practice, we incorporate the anatomy information with four different ratios: 10%, 30%, 50%, and 80% to train the AWCL (fine-grained) models on the pre-training dataset, respectively. Then, we evaluate these trained models on Task II via full fine-tuning. As shown in Fig. 6, we observe that the performance improves consistently with the increasing anatomy ratio. It suggests that using more distinct but anatomically similar samples to compose positive pairs results in better quality representations.

7 Conclusion

In this paper, we presented a new anatomy-aware contrastive learning approach for fetal ultrasound imaging tasks. The proposed approach is able to leverage more positive samples from same or different US videos with the same anatomy category and align well with the anatomical characteristics of ultrasound videos. The feature representative analysis shows our pre-trained representations are more discriminative so as to be better generalized to downstream tasks. Through the reported comparative study, AWCL with fine-grained anatomy information which preserves intra-class difference was more effective than its counterpart. Extensive experimental evaluations demonstrate that our AWCL approach is advantageous to provide useful transferable representations for various downstream clinical tasks, especially for cross-domain generalization. The proposed approach can be potentially applied to other medical imaging modalities where such anatomy information is available.

Acknowledgement

The authors would like to thank Lok Hin Lee, Richard Droste, Yuan Gao and Harshita Sharma for their help with the data preparation. This work is supported by the EPSRC Programme Grant Visual AI (EP/T028572/1), the Project Seebibyte (EP/M013774/1), the ERC Project PULSE (ERC-ADG-2015 694581), the NIH grant U01AA014809, the NIHR Oxford Biomedical Research Centre, and the NVIDIA Corporation with the donation of the GPU.

References

1. Fetal Anomaly Screen Programme Handbook. NHS Screening Programmes, London (2015)
2. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M.: Big self-supervised models advance medical image classification. arXiv:2101.05224 (2021)
3. Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D.: Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In: MICCAI (2019)
4. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11), 2204–2215 (2017)
5. Cai, Y., Droste, R., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. Medical Image Analysis **65**, 101762 (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)
7. Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X.: USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In: MICCAI. pp. 627–637 (2021)
8. Droste, R., Cai, Y., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Ultrasound image representation learning by modeling sonographer visual attention. In: Information Processing in Medical Imaging (IPMI) (2019)
9. Drukker, L., et al.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Scientific Reports (2021)
10. Haghighi, F., Hosseinzadeh Taher, M.R., Zhou, Z., Gotway, M.B., Liang, J.: Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: MICCAI. pp. 137–147 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
13. Hosseinzadeh Taher, M.R., Haghighi, F., Feng, R., Gotway, M.B., Liang, J.: A systematic benchmarking analysis of transfer learning for medical image analysis. In: MICCAI Workshops (2021)

14. Hu, S.Y., Wang, S., Weng, W.H., Wang, J., Wang, X., Ozturk, A., Li, Q., Kumar, V., Samir, A.E.: Self-supervised pretraining with DICOM metadata in ultrasound imaging. In: Proceedings of the 5th Machine Learning for Healthcare Conference. pp. 732–749 (2020)
15. Islam, A., Chen, C.F.R., Panda, R., Karlinsky, L., Radke, R., Feris, R.: A broad study on the transferability of visual representations with contrastive learning. In: IEEE International Conference on Computer Vision (ICCV). pp. 8845–8855 (2021)
16. Jiao, J., Cai, Y., Alsharid, M., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-supervised contrastive video-speech representation learning for ultrasound. In: MICCAI (2020)
17. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-supervised representation learning for ultrasound video. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1847–1850. IEEE (2020)
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
19. Kiyasseh, D., Zhu, T., Clifton, D.A.: CLOCS: contrastive learning of cardiac signals across space, time, and patients. In: International Conference on Machine Learning (ICML). vol. 139, pp. 5606–5615 (2021)
20. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9** (2008)
21. Paszke, et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019)
22. Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D.: Attention-gated networks for improving ultrasound scan plane detection. In: International Conference on Medical Imaging with Deep Learning (MIDL) (2018)
23. Sharma, H., Drukker, L., Chatelain, P., Droste, R., Papageorghiou, A., Noble, J.: Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos. Medical Image Analysis **69**, 101973–101973 (2021)
24. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. In: Medical Imaging with Deep Learning (MIDL) (2021)
25. Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A.Y., Rajpurkar, P.: Medaug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. In: Machine Learning for Healthcare Conference. vol. 149, pp. 755–769 (2021)
26. Zhou, H.Y., Yu, S., Bian, C., Hu, Y., Ma, K., Zheng, Y.: Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In: MICCAI (2020)
27. Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: Generic autodidactic models for 3D medical image analysis. In: MICCAI. pp. 384–393 (2019)
28. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3D medical images by playing a rubik's cube. In: MICCAI (2019)