Fetal Ultrasound Video Representation Learning using Contrastive Rubik's Cube Recovery

Kangning Zhang[1], Jianbo Jiao[1,2], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
kangning.zhang@eng.ox.ac.uk
[2] School of Computer Science, University of Birmingham, UK

Abstract. Contrastive learning (CL), which relies on the contrast between positive and negative pairs, has become the leading paradigm in self-supervised learning. In this paper, we propose a self-supervised learning framework, the feature-level Contrastive Rubik's Cube Recovery (CRCR). CRCR creates contrastive sub-cube pairs from ultrasound video, which capture local spatio-temporal ultrasound features, unlike traditional CL methods which are spatial and work at the global frame level. This approach learns a representation with both intra- and inter-feature contrast to provide strong local feature discrimination. The proposed method is validated on two fetal ultrasound video tasks. Extensive experiments demonstrate that our approach is effective for learning representations that transfer to both in-domain (second-trimester) and cross-domain (first-trimester) clinical downstream classification tasks. In particular, CRCR outperforms four state-of-the-art contrastive learning-based methods on the in-domain task by 3.8%, 2.0%, 1.9% and 1.1%, with each improvement being statistically significant. Code is available at: https://github.com/kangning-zhang/CRCR.

Keywords: Ultrasound · Self-supervised · Contrastive Learning.

1 Introduction

Ultrasound (US), due to its safety and portability, has become one of the most common medical imaging techniques for fetal health monitoring in prenatal care [1][20]. However, human annotation of fetal US images and videos could be expensive, and sometimes infeasible to obtain. Self-supervised learning (SSL) has been applied to US analysis to achieve promising results in US diagnostic tasks using a small amount of labelled data [18]. Most prior works focus on pretext tasks applied to US images, aiming to learn representations through spatial transformations [3][7][12][27]. As US scanning may include a video recording of the US scan, some recent works explore SSL for the entire video instead of video frames, to learn both spatial and temporal representations. Jiao et al. propose a joint reasoning approach to learn representations from both order correction and geometric transformation [10]. As contrastive learning (CL) has become one of the leading paradigms of SSL [4], Chen et al. propose the US semi-supervised contrastive learning (USCL) method [5] and Zhang et al. design the hierarchical contrastive (HiCo) learning method [26] for US video, which currently provides state-of-the-art performance.

Most existing US video pre-training methods generate contrastive pairs using video-level data augmentations [7][5][26]. Two main types of augmentations are spatio-temporal transformations (e.g. cropping, shuffling) and colour transformations (e.g. solarization) [17]. Normally, augmented views from the same video are referred to as positive pairs, and samples from different videos are referred to as negative pairs. CL learns global representation by relying on the representation invariant of positive pairs [4]. However, given the fact that fetal US videos often share a global spatial pattern with significant local variations, we are motivated to explore the use of local contrastive pairs generated from sub-cubes of US videos for local representation learning.

In this paper, we address this issue by proposing a SSL framework Feature-level Contrastive Rubik's Cube Recovery (CRCR) for US video representation learning. We introduce Rubik's cube recovery (RCR) [29] and cube reconstruction [13], which are pretext tasks designed to learn spatio-temporal context by image restoration, as effective tools to provide strong spatio-temporal distortions and create contrastive pairs from sub-cubes of US video. Unlike recent methods DiRA [8] and Swin UNETR [19], which leverage other pretext-tasks to CL frameworks by directly combining the training objectives of each pretext task. Our method, motivated by [22], provides a novel approach to generate both inter- and intra-feature contrastive pairs based on the introduced pretext tasks. Here, inter-feature pairs include sub-cubes from distinct US videos, and intra-feature pairs include distinct distorted sub-cubes from the same US video. We hypothesise that our approach could provide stronger local discrimination and enhance local representation learning.

In summary, our main contributions are as follows: 1) We propose a SSL framework, called feature-level Contrastive Rubik's Cube Recovery (CRCR) for fetal ultrasound video, which is customized to combine contrastive learning with Rubik's cube recovery and cube reconstruction; 2) We introduce an approach to generate local contrastive pairs from sub-cubes of US video, which facilitate discriminative and consistent local representation learning; 3) We empirically compare the effect of using feature-level pretext tasks and stronger feature extractor (i.e. 3D Swin Transformer) to enhance feature learning; 4) The proposed method CRCR consistently outperforms several existing SSL methods on both in-domain and cross-domain US clinical downstream tasks, showing its effectiveness and generalisability.

2 Methods

2.1 CRCR Framework

Suppose $x_i \in \mathcal{X} \subset \mathbb{R}^{1 \times H \times W \times K}$ is a video clip in an US video dataset, where 1 represents the grey-scale property of US and H, W and K denotes the height, width, number of frames in the video clip, respectively.

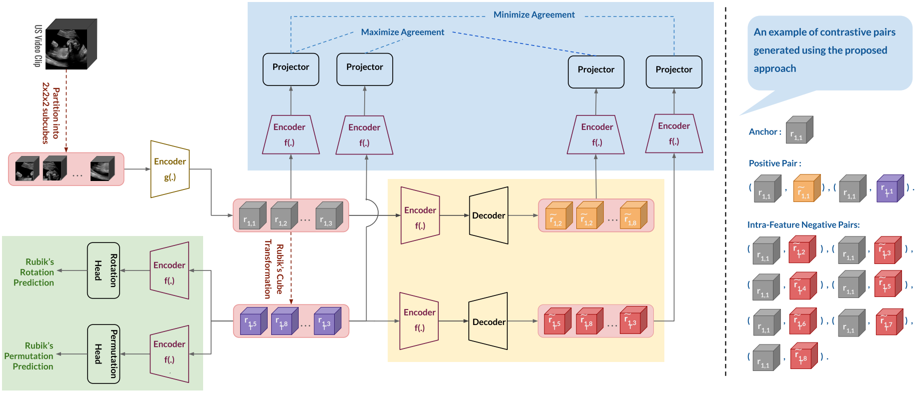

Fig. 1: The pipeline of the proposed Contrastive Rubik's Cube Recovery (CRCR) framework, consisting of three pretext tasks: contrastive learning (the blue block), cube reconstruction (the yellow block), and Rubik's Cube recovery (the green block). An example of our proposed local contrastive pair generation approach is demonstrated with a given anchor.

The overall framework of our method is illustrated in Fig. 1, which consists of two encoders $g(.)$ and $f(.)$, one projection head $h(.)$ for contrastive learning, two MLP head $r(.)$ and $p(.)$ for Rubik's cube recovery task and one decoder $d(.)$ for the cube reconstruction task.

Different from the normal CL paradigm [4], CRCR introduces two novel designs for effective local representation learning. Firstly, we perform pretext tasks on the feature level instead of the video level, and secondly, we proposed an approach to generate local contrastive pairs from sub-cubes of US video. Since the quality of US videos is always affected by the extensive presence of speckle noises and acoustic shaded [11], whereas low-level information (i.e. boundaries) and local noise could be discarded at the encoder features. It is worth considering operating pretext tasks at the feature level instead of the video level to learn stronger and meaningful representations. As shown in Fig. 1, an input US video clip x_i is firstly partitioned into $2 \times 2 \times 2$ sub-cubes $\{x_{i,j}\}_{j=1}^{8}$, following the partition of Rubik's cube. The sub-cubes of US video along with its position embedding are then sent into the encoder $g(.)$ to get sub-cubes of US features $\{r_{i,j}\}_{j=1}^{8}$. Distortions (Rubik's cube transformation and reconstruction) are operated on the obtained features. Local contrastive pairs are generated from sub-cube of US features with the designed distortions, as in Sec. 2.2.

2.2 Training Objectives

Rubik's cube recovery. RCR, which includes both rotation and permutation, are operated on sub-cubes $\{r_{i,j}\}_{j=1}^{8}$. Here, we define the set of permutations $\mathcal{P} = \{p_1, p_2, ..., p_K\}$ to have the largest Hamming distance for sub-cubes shuffling and the set of rotations $\mathcal{R} = \{r_1, r_2, ..., r_M\}$ to ensure either a horizontal or vertical flip for each sub-cube, maximising the distortion. $P_k \sim \mathcal{P}$ and $R_{m,j} \sim \mathcal{R}$ are sampled. The transformed sub-cubes are denoted as $\{r_{i,j}^T\}_{j=1}^8$, with position embedding updated to aligned with P_k. The predicted results l_j and $\{g_{m,j}\}_{j=1}^8$ are obtained from the rotation and permutation heads, respectively. RCR loss is calculated as the sum of rotation and permutation loss as follows:

$$L_{RCR} = -\{\sum_{k=1}^{K} p_k \log l_k + \sum_{j=1}^{8} \sum_{m=1}^{M} r_{m,j} \log g_{m,j}\}. \tag{1}$$

Cube reconstruction. Cube reconstruction is operated on the sub-cubes $\{r_{i,j}\}_{j=1}^8$ and $\{r_{i,j}^T\}_{j=1}^8$, with the latter's position embedding updated. The obtained reconstructions are denoted as $\{\tilde{r_{i,j}}\}_{i=1}^8$ and $\{\tilde{r_{i,j}}^T\}_{i=1}^8$, with the latter's position embedding updated. The reconstruction loss is calculated as follows:

$$L_{Reconst.} = \alpha_1 \times \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}, r_{i,j}) + \alpha_2 \times \sum_{j=1}^{8} MSE(\tilde{r_{i,j}}^T, r_{i,j}^T). \tag{2}$$

Contrastive learning. With the assumption that US videos share a global spatial pattern with local divergence, we propose generating local contrastive pairs using sub-cubes of US videos, instead of global pairs using the entire video.

Normally, CL considers different US videos as negative pairs, which might result in the high similarity between negative pairs (i.e. videos from the same scan or performing the same measurement task) and potentially mislead representation learning [5]. The proposed approach is designed to generate two sets of strongly discriminative negative samples $I^- = \{I_{intra}^-, I_{inter}^-\}$ by introducing the aforementioned pretext tasks as effective distortion tools. We assume that rotating and shuffling the cube would cause severe spatial and temporal distortion, resulting in the loss of both spatial and temporal information and leading to poor-quality reconstruction. A Local positive samples set, I^+, is generated with appropriate similarities. For a given anchor sub-cube $r_{i,j}$, the local contrastive pairs generated from the proposed approach consist of:

- **Positive sample** $I^+ = \{r_{i,j}^T, \tilde{r_{i,j}}\}$: rotated and reconstructed views of anchor
- **Intra-feature negative samples** $I_{intra}^- = \{\tilde{r_{i,j}}^T, ..., \tilde{r_{i,j}}^T\}$: distorted sub-cubes from the remaining sub-cubes within the same video
- **Inter-feature negative samples** $I_{inter}^- = \{\tilde{r_{k,l}}^T\}_{k=1 \neq i, l=1}^{N \quad 8}$: distorted sub-cubes from different videos

Those two sets of negative samples enable the model to learn representations from different perspectives. Inter-feature negative pairs enhance instance discrimination, while intra-feature negative pairs provide local contextual information. Referred to [4][25], we adapt NT-Xent for our contrastive loss function, which is calculated as follows:

$$
\begin{aligned}
L_{CLR} &= -\sum_{j=1}^{8} \sum_{i^+ \in I^+} \log \frac{\exp(\varphi(r_{i,j}, i^+)/\tau)}{\sum_{i^- \in I^-} \exp(\varphi(r_{i,j}, i^-)/\tau)} \\
&= -\sum_{j=1}^{8} \sum_{i^+ \in I^+} \{\varphi(r_{i,j}, i^+)/\tau - \log \sum_{i^- \in I^-} \exp(\varphi(r_{i,j}, i^-)/\tau)\}
\end{aligned}
\tag{3}
$$

where τ and $\varphi(.)$ denote the temperature parameter and the pairwise cosine similarity function, respectively.

Overall loss function. The overall learning target is a weighted combination of Rubik's cube recovery loss, reconstruction loss, and contrastive loss,

$$L_{CRCR} = \alpha \times L_{Rubik} + \beta \times L_{Reconst.} + \eta \times L_{CLR}. \tag{4}$$

A grid-search hyperparameter optimization was performed which estimated the optimal values of $\alpha = \eta = 1, \beta = 0.5$.

2.3 Stronger Feature Extractor

We propose to replace the traditional convolutional neural network with a stronger feature extractor. 3D Swin Transformer [14][23], as an effective transformer-based backbone, is considered. In 3D patch partition module, we take a sub-cube of size $\frac{T}{2} \times \frac{H}{2} \times \frac{W}{2}$ as the basic processing unit, which align with the partition of a Rubik's cube. A linear embedding is then applied to project the feature into a C-dimensional space and a positional embedding, which would be updated with the applied permutations, is added to retain sub-cube's positional information. The subsequent training stages follow the implementation described in [23].

2.4 Implementation Details

Both 3D ResNet-18 [9] and 3D Swin Transformers [23] are utilized as the backbone networks for the proposed self-supervised learning method. ResNet-18 is chosen based on prior work [5][28], while Swin Transformer is chosen for its strong feature extraction capability. We return each sub-cube to its embedded position before passing it to 3D ResNet encoders to include positional information.

For SSL pretraining, we employ an AdamW optimizer [15] to be consistent with [19]. Both networks are trained with a momentum of 0.9, a warm-up cosine scheduler of 500 iterations and a mini-batch of 32 for 40 epochs. After parameter tuning, an initial learning rate of 1×10^{-3} is used with decay of 0.1 for every 25K iterations for 3D ResNet-18 and an initial learning rate of 1×10^{-3} is used with decay of 0.01 for every 45K iteration for 3D Swin Transformer. All models are implemented with PyTorch [16], with our methods taking around 180 hours to run on a single NVIDIA Titan V GPU.

For transfer learning, we fine-tune the combined encoder $f(g(.))$ along with an attached classifier head. For the standard plane detection task, networks are trained with SGD optimizer with momentum of 0.9 and mini-batch of 16 for 70 epochs. An initial learning rate of 0.01 is used with 0.1 decay at epochs 30 and 55. For first-trimester anatomies recognition, networks are trained with SGD optimizer with momentum 0.9 and mini-batch 16 for 200 epochs. An initial learning rate of 0.1 is used with a decay of 0.1 at epochs 150.

3 Results

3.1 Ultrasound Data

Our experiments are based on a large-scale fetal Ultrasound (US) video dataset [6]. Full-length routine fetal ultrasound videos are recorded and sampled at the rate of 30 Hz using a commercial Voluson E8 version BT18 ultrasound machine. We consider a subset of the entire dataset for pre-training, in which only scan recordings of the second trimester (gestational age of 18–22 weeks) are considered. The pre-training dataset consists of a total number of 70,661 video clips (each of length 32, with 2,261,179 frames in total) from 719 US scan recordings. 135 second-trimester scans are selected for three-fold cross-validation on the standard plane detection task, which consists of 15,384 labelled frames. A subset of first-trimester scans is used for cross-domain anatomy recognition task, which consists of 55,871 frames with 5 anatomy categories. All frames were central cropped to remove the fan shape and resized to 224 × 224 pixels.

3.2 Transfer Learning on Standard Plane Detection

Task description. We evaluate the pre-trained representation by transferring it to the in-domain second-trimester standard plane detection task. Similar to [2], 14 classes are considered, which include four cardiac views, three-vessel and trachea (3VT), four-chamber (4CH), right ventricular outflow tract (RVOT), and left ventricular outflow tract (LVOT), two brain views, transventricular (BrainTv.) and transcerebellum (BrainTc.), two spine views, coronal (SpineCor.) and sagittal (SpineSag.), abdominal, femur, kidneys, lips, profile and background class. Precision, recall and F1-scores as used as the evaluation metrics.

Results. Table 1 shows a quantitative comparison of fine-tuning performance on the standard plane detection task. The table indicates that CRCR generally outperforms four state-of-the-art CL-based methods, i.e. SimCLR [4], DiRA [8], USCL [5], and HiCo [26], by 3.8%, 1.9%, 2.0% and 1.1%, respectively, in F-1 score with the 3D ResNet backbone. Additionally, CRCR improves the performance of SimCLR and Swin UNETR by 3.7% and 1.9%, when using 3D Swin Transformer as the backbone. These improvements are statistically significant using the Wilcoxon signed-rank test [21], validating the effectiveness of our proposed method. In particular, CRCR performs better with 3D Swin Transformer

Table 1: Quantitative comparison of downstream task performance (mean±std.[%]) on second-trimester standard plane detection task and first-trimester anatomy recognition task. *Rand. Init.* indicates the 3D ResNet18 trained from scratch, and ‡ denotes the use of 3D Swin Transformer as backbone network. The best results for each backbone network are marked in **bold**. P-values are calculated between our CRCR results and the previous top-1 result for each backbone network. Any $p < 0.05$ represents a statistically significant improvement and is highlighted in green.

	Standard Plane Detection			Anatomy Recognition		
Pretrain Method	Precision	Recall	F1	Precision	Recall	F1
Rand. Init.	69.3±1.6	58.8±3.1	59.1±3.1	80.8±3.2	78.4±0.7	81.1±0.5
RCR [29]	70.0±0.8	66.3±3.5	66.4±2.3	89.2±1.4	88.6±1.6	89.5±1.7
SimCLR [4]	70.9±0.5	68.9±1.7	68.7±1.2	95.5±0.5	94.7±0.3	94.8±0.8
DiRA [8]	72.0±2.5	70.4±2.3	70.6±2.3	95.7±1.4	95.3±1.3	95.1±1.2
USCL [5]	71.6±1.1	70.2±1.5	70.5±0.9	95.9±0.7	95.2±0.8	95.3±1.5
HiCo [26]	72.3±1.7	71.7±1.9	71.4±2.0	96.3±0.3	95.7±0.7	95.8±0.9
CRCR (ours)	**73.1 ±2.3**	**72.6±2.7**	**72.5±2.2**	**96.8±1.2**	**96.1±1.4**	**96.2±1.8**
P-value	0.048	0.021	0.010	0.018	0.027	0.035
SimCLR[4]‡	72.8±0.8	70.7±1.2	70.7±1.6	95.9±1.3	95.3±0.8	95.1±1.5
SwinUNETR[19]‡	73.6±1.2	73.2±3.5	72.5±2.1	96.7±2.7	96.9±2.4	96.5±2.3
CRCR (ours)‡	**74.9±2.5**	**74.8±3.0**	**74.4±2.6**	**97.1±2.3**	**97.1±1.6**	**97.2±1.3**
P-value	0.003	0.001	<0.001	0.042	0.062	0.009

as the backbone network, demonstrating the benefit of utilising a stronger feature extractor to enhance feature learning. Supp. Table 1 presents the F1-score for each class, in which CRCR owns the best performance in the majority of classes. The improvement is particularly notable for *cardiac* views, which share a global heart perspective but each has a local focus on specific structures, making them difficult to distinguish even for experts. This finding is in line with our assumption that by contrasting local contrastive pairs, the proposed method can learn semantically meaningful information from these local areas.

3.3 Transfer Learning on First-Trimester Anatomy Recognition

Task description. We explore the generalisability of the pre-trained representation to a cross-domain first-trimester anatomy recognition task. Similar to [24], five key anatomy categories are considered, which include crown rump length (CRL), nuchal translucency (NT), biparietal diameter (BPD), 3D-mode (3D) and other (BK) for first-trimester fetal biometry measurements.

Results. Table 1 demonstrates quantitative results of fine-tuning performance on the first-trimester anatomy recognition task. From the results, we observe that our proposed method achieves the best performance among all the compared methods with both 3D ResNet and 3D Swin Transformer as backbone networks. Most improvements are statistically significant with paired t-tests. This task of

Table 2: Ablation studies on the effectiveness of each self-supervised objective and the effectiveness of performing pretext tasks at the feature level. Experiments are fine-tuned using 3D ResNet for the standard plane detection task.

			Feature-level			Video-level		
L_{CLR}	L_{Rubik}	L_{Rec}	Precision	Recall	F1-score	Precision	Recall	F1-score
✓			71.2	69.4	69.5	70.9	68.9	68.7
	✓		70.4	66.7	66.4	70.0	66.3	66.4
		✓	70.3	66.1	65.8	69.8	64.2	64.5
	✓	✓	70.9	69.0	68.5	70.6	68.5	68.1
✓	✓		72.3	71.1	70.9	71.4	69.7	69.8
✓		✓	71.8	70.5	70.3	71.0	69.5	69.1
✓	✓	✓	**73.1**	**72.6**	**72.0**	**72.4**	**71.1**	**71.3**

anatomical recognition could be challenging due to the small fetal size in the first trimester. While the compared CL-based methods focus on global representation learning, CRCR leans local patterns through local contrastive pairs, which could be valuable when the clinical region-of-interest is small. Among the compared methods, USCL and HiCo, which are designed specifically for US videos, perform better than those designed for general computer vision. Supp. Table 2 shows the F1-score for each anatomy category, in which CRCR achieves the highest F1 scores in all categories. This finding demonstrates the effectiveness and generalisability of our proposed method CRCR on first-trimester US video.

3.4 Ablation Study

Efficacy of self-supervised objectives. We perform an empirical study on pre-training with different combinations of self-supervised objectives used in CRCR loss. Results are shown in Table 2. An improvement could be seen by adding pretext tasks, with the combination of contrastive learning and Rubik's cube recovery task as the best-performing pairing. These results indicate that three selected tasks harmonize with each other and the proposed collaborative learning methods enhance representation learning and downstream performance.

Efficacy of feature-level pretext task. We investigate how performing pretext tasks at the feature-level impacts representation learning compared to the video level. As illustrated in Table 2, performing pretext tasks at video level degrades the performance of the standard plane detection task to feature level. This aligns with our hypothesis that performing pretext tasks at the feature level enables the model to be less sensitive to superficial changes and focus more on informative regions, as local noise and irrelevant features could be discarded through encoder optimization.

4 Conclusion

In this paper, we present a novel self-supervised learning method named feature-level Contrastive Rubik's Cube Recovery (CRCR) for local representation learning of fetal US video. The proposed method leverages the advantages of contrastive learning with Rubik's cube recovery task and cube reconstruction task. A local contrastive pair generation approach is introduced to contrast sub-cube pairs from US video. Through extensive experiments, it is demonstrated that CRCR achieves state-of-the-art performance on both second-trimester standard plane detection task and first-trimester anatomy recognition task and significantly improves the quality of learnt representation in pre-training for both in-domain and cross-domain downstream tasks. In the future, the proposed approach can be potentially applied to other medical imaging modalities.

Acknowledgement

The authors would like to thank the support from the ERC Project PULSE ERC-ADG-2015 694581 and the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1). Jianbo Jiao is supported by the Royal Society Short Industry Fellowship (SIF\R1\231009).

References

1. Abramowicz, J.S.: Benefits and risks of ultrasound in pregnancy. In: Seminars in perinatology. vol. 37, pp. 295–300. Elsevier (2013)
2. 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)
3. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical image analysis **58**, 101539 (2019)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
5. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 627–637. Springer (2021)
6. Drukker, L., Sharma, H., Droste, R., Alsharid, M., Chatelain, P., Noble, J.A., Papageorghiou, A.T.: Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video. Scientific Reports **11**(1), 14109 (2021)
7. Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Anatomy-aware contrastive representation learning for fetal ultrasound. In: European Conference on Computer Vision. pp. 422–436. Springer (2022)

8. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20824–20834 (2022)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. 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)
11. Li, H., Fang, J., Liu, S., Liang, X., Yang, X., Mai, Z., Van, M.T., Wang, T., Chen, Z., Ni, D.: Cr-unet: A composite network for ovary and follicle segmentation in ultrasound images. IEEE journal of biomedical and health informatics **24**(4), 974–983 (2019)
12. Liu, H., Liu, J., Hou, S., Tao, T., Han, J.: Perception consistency ultrasound image super-resolution via self-supervised cyclegan. Neural Computing and Applications pp. 1–11 (2023)
13. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering **35**(1), 857–876 (2021)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
17. Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K.: Self-supervised learning: A succinct review. Archives of Computational Methods in Engineering **30**(4), 2761–2775 (2023)
18. Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Computer Science **8**, e1045 (2022)
19. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
20. Whitworth, M., Bricker, L., Mullan, C.: Ultrasound for fetal assessment in early pregnancy. Cochrane database of systematic reviews (7) (2015)
21. Woolson, R.F.: Wilcoxon signed-rank test. Encyclopedia of Biostatistics **8** (2005)
22. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8392–8401 (2021)
23. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. arXiv preprint arXiv:2304.06906 (2023)
24. Yasrab, R., Fu, Z., Zhao, H., Lee, L.H., Sharma, H., Drukker, L., Papageorgiou, A.T., Noble, J.A.: A machine learning method for automated description and workflow analysis of first trimester ultrasound scans. IEEE Transactions on Medical Imaging **42**(5), 1301–1313 (2022)

25. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y.: Graph contrastive learning with augmentations. Advances in neural information processing systems **33**, 5812–5823 (2020)
26. Zhang, C., Chen, Y., Liu, L., Liu, Q., Zhou, X.: Hico: hierarchical contrastive learning for ultrasound video model pretraining. In: Proceedings of the Asian Conference on Computer Vision. pp. 229–246 (2022)
27. Zhang, J., He, Q., Xiao, Y., Zheng, H., Wang, C., Luo, J.: Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network. Medical Image Analysis **70**, 102018 (2021)
28. Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y.: Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. Medical image analysis **64**, 101746 (2020)
29. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 420–428. Springer (2019)