2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340504

The Effectiveness of Self-supervised Pre-training for Multi-modal Endometriosis Classification*†

David Butler[1], Hu Wang[1], Yuan Zhang[1], Minh-Son To[2], George Condous[3], Mathew Leonardi[4], Steven Knox[5], Jodie Avery[6], M Louise Hull[6], Gustavo Carneiro.[7]

***Abstract*— Endometriosis is a debilitating condition affecting 5% to 10% of the women worldwide, where early detection and treatment are the best tools to manage the condition. Early detection can be done via surgery, but multi-modal medical imaging is preferable given the simpler and faster process. However, imaging-based endometriosis diagnosis is challenging as 1) there are few capable clinicians; and 2) it is characterised by small lesions unconfined to a specific location. These two issues challenge the development of endometriosis classifiers as the training datasets tend to be small and contain difficult samples, which leads to overfitting. Hence, it is important to consider generalisation techniques to mitigate this problem, particularly self-supervised pre-training methods that have shown outstanding results in computer vision and natural language processing applications. The main goal of this paper is to study the effectiveness of modern self-supervised pre-training techniques to overcome the two issues mentioned above for the classification of endometriosis from multi-modal imaging data. We also introduce a new masking image modelling self-supervised pre-training method that works with 3D multi-modal medical imaging. Furthermore, to the best of our knowledge, this paper presents the first endometriosis classifier, fine-tuned from the pre-trained model above, which works with multi-modal (i.e., T1 and T2) magnetic resonance imaging (MRI) data. Our results show that self-supervised pre-training improves endometriosis classification by as much as 31%, when compared with classifiers trained from scratch.**

***Index Terms*— Self-supervision, Multi-modal learning, MRI, Endometriosis**

I. Introduction

Endometriosis is a gynaecological disorder characterised by the abnormal growth of the uterine lining on the exterior of the uterus and surrounding organs [1], which affects approximately 5-10% of women [2]. There are currently no known preventative methods, however, early detection and treatment can help manage the symptomatic impact and reduce the further progression of the disease [3]. One of the most reliable methods for diagnosing endometriosis is based on surgery [4]. While having good accuracy, surgical procedures are invasive, have the possibility of complications [4], and can be expensive [5], which motivate the interest in an alternative way to diagnose endometriosis through non-invasive analysis of multi-modal (e.g., T1, T2) images derived from magnetic resonance imaging (MRI) [4]. However, the diagnosis of endometriosis from medical imaging can be challenging due to the small size of lesions that can be widespread throughout the abdominal cavity, and the limited availability of clinicians who can successfully diagnose endometriosis from MRI.

These two issues result in small training sets with challenging samples to train MRI-based endometriosis classifiers. In general, data augmentation and weight decay are enough to mitigate similar training issues, but the two issues mentioned above are too challenging for these solutions, as we will present in this paper. Therefore, the main research question we aim to answer with this paper is if modern self-supervised pre-training techniques can help with the generalisation of endometriosis classifiers using multi-modal MRI data, assuming that we have available a large dataset of unlabelled female pelvic multi-modal MRI data.

In this paper, we study several self-supervised pre-training methods, which rely on unlabelled female pelvic multi-modal MRI data, to allow the effective generalisation of an endometriosis classifier that processes multi-modal MRI images. The main contributions of this paper are:

- To the best of our knowledge, this is the first deep learning classifier that can automatically diagnose endometriosis from 3D multi-modal MRI data;
- A new masking image modelling self-supervised pre-training method based on a masked auto-encoder [6] applied to 3D multi-modal MRI data; and
- A comprehensive study of several types of self-supervised pre-training methods, based on siamese networks [7] and masked image modelling [8], trained from a large unlabelled female pelvic multi-modal MRI dataset to enable an effective generalisation of the endometriosis classifier;

Our experiments show that recently proposed self-supervised pre-training methods, particularly the ones based on masking image modelling, can be successfully used to overcome the limitations caused by the small training sets and challenging classification of endometriosis from multi-modal MRI. Results show that self-supervised pre-training improves endometriosis classification by 5% to 31%.

*This work received funding from the Australian Government through the Medical Research Futures Fund: Primary Health Care Research Data Infrastructure Grant 2020 and from Endometriosis Australia.

†This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Southern Adelaide Clinical Human Research Ethics Committee of the Southern Adelaide Local Health Network (Date 22 Sept 2021/No. 111.20).

[1]Australian Institute for Machine Learning, University of Adelaide, Australia
[2]Flinders Health and Medical Research Institute, Flinders University, Australia
[3]Omnigynaecare, Sydney, Australia
[4]McMaster University, Canada
[5]Benson Radiology, Adelaide, Australia
[6]Robinson Research Institute, University of Adelaide, Australia
[7]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

II. Related Work

A. Self-supervised Learning

The mitigation of problems associated with small training sets or hard training problems usually involves the use of regularisation techniques to avoid overfitting. The most common techniques are based on weight decay or data augmentation, but they only work for simple cases. More difficult cases are likely to require the more effective self-supervised pre-training methods, which are designed to improve the generalisation of models with the use of unlabelled data. More specifically, these methods generate their own supervisory signal from unlabelled data to learn effective embeddings for downstream tasks [9], alleviating the need to learn such basic embeddings from the smaller labelled datasets. Current popular methods include the use of siamese networks and masked image modelling.

Siamese networks rely on two similar networks that encode pairs of images [10], usually representing different views of the same image, with similar embeddings. These methods use various techniques to avoid the collapse to the trivial solution, where all images have the same representation. Some repel negative pairs in the embedding space [11], while others use momentum encoders [7] or clustering [12]. **Masking image modelling** (MIM) attempts to learn representations by "filling in the blanks" [6] of a reconstructed image. In particular, the model is trained to reconstruct the input image using only a fraction of this image. By attempting to predict the hidden parts, the model also learns to extract underlying important features.

The MIM methods generally rely on vision transformer models [8], such as the **vision transformer (ViT)**. The critical feature of ViT is self-attention, which unlike convolutional neural networks (CNNs), allows for the dynamic combination of distant features in a single step. This self-attention, however, also increases the solution space of ViT, leading the field to propose solutions that allow vision transformers to perform effectively even with small datasets, but with restrictions on image shape [13]. Initial methods [14] used a separate tokeniser to create tokens for the encoder to predict. More recently, masked autoencoder (MAE) [6] has shown that a separate tokeniser is unnecessary, and reconstruction can be done using an autoencoder structure. MIM has also been shown to work well with 3D inputs, such as video or volumes [15], [16].

B. Multi-modal Learning from 3D Medical Imaging

Multi-modal learning utilises information from multiple data modalities when performing a task [17]. Modalities can have different formats, such as text, image, and audio, or different semantics, as in our case, with T1 and T2 weighted MRI volumes. To reconcile modalities, related representations are needed [18]. This can be done by learning a joint representation through fusion, where both modalities share a space, coordinated representations, where the representations have linking constraints, or through translation, where one modality is translated into another [17].

Transformer-based multi-modal methods have been studied because attention is adaptable to different multi-modal techniques. Self-attention dynamically weights inter and intra-modality features simultaneously, such as with MultiMAE [19], which extends MAE to multiple modalities by concatenating the inputs for early fusion. Whether MultiMAE can perform well on 3D or unaligned inputs has not been investigated. Alternatively, cross-modal attention uses query, key, and value combinations from separate modalities [20], [21]. This encourages extracting features with shared semantics while keeping the spaces separate.

Recent work in **3D medical image** analysis has focused on applying vision transformers [22], [23]. In addition, pre-training techniques have been investigated [22], with the application of MIM [16], including MAE [23], showing promising results. However, little work has been done to investigate the use of 3D multi-modal data in MIM.

C. Endometriosis Classification

We are not aware of classifiers that work for classifying endometriosis from 3D multi-modal MRI data. Few studies have investigated the use of shallow machine learning techniques to predict the presence or extent of endometriosis based on the presence of symptoms or signs [24], [25]. As for harnessing imaging data, there have been few studies on classifying endometriosis or other endometrial-related diseases. Related work has been done on detecting endometrial cancer from 2D images [26], while Maicas et al. [27] use temporal residual networks to detect pouch-of-Douglas (POD) obliteration. The most related work is the MRI and ultrasound endometriosis classifier by Yang et al. [28], but they rely on an unrealistic setup based on the availability of paired MRI and ultrasound data from the same patient, which rarely happens in practice, and differently from us, they use single-modal MRI data.

III. Methods

Let $\mathcal{D} = \{(\mathbf{x}_i^{T1}, \mathbf{x}_i^{T2}, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$ represent the endometriosis MRI dataset, where $\mathbf{x}^{T1}, \mathbf{x}^{T2} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times D}$ are the T1 and T2 MRI images with height, width and depth denoted by H, W and D, and $\mathbf{y} \in \{0, 1\}$ is the endometriosis classification label. For the self-supervised pre-training, we have $\mathcal{D}_U = \{(\mathbf{x}_i^{T1}, \mathbf{x}_i^{T2})\}_{i=1}^{|\mathcal{D}_U|}$, with $|\mathcal{D}_U| >> |\mathcal{D}|$. The endometriosis classifier is denoted by $f_\theta : \mathcal{X} \rightarrow \Delta$, where $\Delta \subset [0, 1]$ is the probability simplex. The early-fusion multi-modal classifier is denoted by $f_\theta^M : \mathcal{X} \times \mathcal{X} \rightarrow \Delta$, while the late fusion classifier uses two separate $f_\theta(.)$ that are pooled by a final fully connected layer, as shown in Fig. 1.

A. Model Backbones

All experiments in this paper are based on three backbones for the classifier $f_\theta(.)$: 3D ResNet18, 3D ViT, and 3D MultiViT. We select 3D ResNet18 [29] as one of the backbones since it performs well on many classification tasks. The pre-training for 3D ResNet18 is based on BYOL [7], as explained below in Sec. III-B, which relies on the embedding produced by the final average pooling layer that has 512 dimensions.

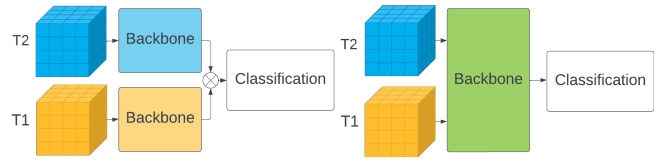

Fig. 1. Late Fusion with ResNet/ViT (Left) vs. Early Fusion with MultiViT (Right).

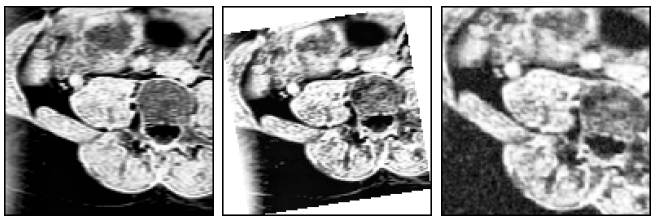

Fig. 2. Augmented views for BYOL. Original (Left), View 1 (Middle), View 2 (Right).

Then, the 3D ResNet18 is fine-tuned by adding a final fully-connected layer. A multi-modal architecture is formed by late fusion, where the outputs of two separate ResNet backbones are concatenated, creating a feature with 1024 dimensions, as shown in Fig. 1.

The vision transformer models have been selected because of their excellent performance in many benchmarks [8]. We use the base variant of ViT, called ViT-B, as a starting point for our modifications. As our task requires 3D inputs, we modify ViT's positional embedding by splitting it into three sections, one for each axis, instead of the usual two. The positional embedding is learnt and initialised with one-dimensional sin-cos embeddings for each axis. This modification restricts the positional embedding size, and thus the embedding size, to multiples of three. We experiment with two different patch sizes, namely: 16^3 from ViT, and our proposed 8^3, which form the 3D ViT-B/16 and 3D ViT-B/8, respectively.

We explore both the early and late fusion strategies for the multi-modal ViT architecture, as shown in Fig. 1. While the late fusion is straightforward to implement, the early fusion is based on applying our ViT modifications to MultiViT [19], forming the 3D MultiViT-B/16 and 3D MultiViT-B/8.

B. Pre-training

The classifier $f_\theta(.)$ can be either trained from scratch using $\mathcal{D}$ or fine-tuned from the pre-trained classifier, where the pre-training uses $\mathcal{D}_U$. We experiment with pre-training methods based on siamese networks and MIM. For the 3D ResNet18 model, we relied only on the siamese network BYOL approach [7] given that it holds state-of-the-art results in many pre-training tasks. As explained in Sec. II-A, BYOL pulls together the embeddings of different views of the same image, formed by data augmentation techniques, such as the ones shown in Fig. 2. On the other hand, for 3D ViT backbone, we explore the MIM masked autoencoder (MAE) [6], and for the 3D MultiViT models, we study the MultiMAE [19] since both MAE and MultiMAE have

Fig. 3. MAE reconstruction of pelvic MRI. Visible (Left), Reconstruction + Visible (Middle), Original (Right).

shown outstanding pre-training results in many benchmarks. MAE and MultiMAE are MIM approaches that aim to learn embeddings by reconstructing the original image when only part of the input image is visible, as shown in Fig. 3.

IV. Experiments

A. Datasets

1) Endometriosis Classification Dataset: Our endometriosis dataset consists of 89 T1/T2 MRI volume pairs cropped to the uterine region with a size of $64 \times 128 \times 128$. Cropping is performed by hand to position the uterus at the front and bottom of the volume to capture the entire pouch-of-douglas (POD) region. Of these, 19 pairs contain complete POD obliteration, and the remaining 70 pairs are free from POD obliteration. Training and evaluation use 5-fold cross-validation, with 71 training pairs and 18 validation pairs in each fold. For pre-processing, we apply adaptive histogram equalisation to improve contrast.

2) Pre-training Dataset: Pre-training used two unlabelled datasets, one for T1 and another for T2. Both contain MRI volumes of female pelvises from various MRI machines at differing resolutions. To standardise, the volumes are re-sampled to 1mm $\times$ 1mm $\times$ 3mm voxels, then either cropped or padded with zeros to a size of $64 \times 128 \times 128$. Adaptive histogram equalisation is also applied to improve contrast. The T1 and T2 datasets contain 5,867 and 8,984 unlabelled volumes, respectively. A joint pre-training dataset is derived from these for training multi-modal approaches by pairing T1 and T2 volumes by patient. Most patients have more than one volume of each modality, so pairings are created for each possible combination. This joint dataset contains a total of 15,048 pairings.

B. Experimental Setup

The classifiers trained from scratch use the AdamW optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.05, base learning rate of $1\text{e}{-3}$, and batch size of 32. The learning rate lr is derived by: $lr = base\_lr \times (batch\_size/256)$. Training lasts for 100 epochs using cosine decay for learning rate scheduling, with a warm-up period of 10 epochs. ViT based models also use 0.75 layer-wise learning rate decay [8]. The data augmentation strategy involves random Gaussian noise and blur, randomly adjusting brightness, contrast, and gamma, and randomly downsampling and re-upsampling to simulate lower resolutions. To extend ViT to three dimensions it is required that the embedding size is divisible by

Pretraining	Backbone	Modality	Avg. AUC (SD)
None	3D ResNet-18	T1	0.752 (0.128)
		T2	0.650 (0.116)
		T1 + T2	0.730 (0.132)
	3D ViT-B/16	T1	0.666 (0.200)
		T2	0.702 (0.165)
		T1 + T2	0.709 (0.143)
	3D ViT-B/8	T1	0.728 (0.164)
		T2	0.713 (0.168)
		T1 + T2	0.646 (0.179)
	3D MultiViT-B/16	T1 + T2	0.742 (0.153)
	3D MultiViT-B/8	T1 + T2	0.661 (0.183)
BYOL	3D ResNet-18	T1	0.745 (0.173)
		T2	0.838 (0.089)
		T1 + T2	0.839 (0.093)
MAE	3D ViT-B/16	T1	0.700 (0.139)
		T2	0.784 (0.124)
		T1 + T2	0.786 (0.118)
	3D ViT-B/8	T1	0.714 (0.092)
		T2	**0.875 (0.086)**
		T1 + T2	0.849 (0.086)
MultiMAE	3D MultiViT-B/16	T1 + T2	0.739 (0.134)
	3D MultiViT-B/8	T1 + T2	0.693 (0.174)

TABLE I
RESULTS FROM 5-FOLD CROSS-VALIDATION.

three, so we round the number of embedded dimensions in the MAE decoder down to 510 from the usual 512.

The BYOL [7], for pre-training the ResNet, has a training that lasts for 100 epochs using Adam as the optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, base learning rate of $1e{-}4$ and batch size of 1024. The learning rate lr is derived by: $lr = base\_lr \times (batch\_size/256)$. The same augmentation strategy used for fine-tuning is applied to derive different views for BYOL pre-training.

The MAE [15], for 3D ViT-B/16 and 3D ViT-B/8, has a training that uses an AdamW optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.05, base learning rate of $1e{-}3$, and batch size of 4. The learning rate lr is derived by: $lr = base\_lr \times (batch\_size/256)$. We rely on a masking rate of 50%. Our setup for MultiMAE is identical to MAE, with one exception. MAE sets the number of visible patches as a percentage of the total number of patches, while MultiMAE uses an absolute value. A masking rate of 50% for MAE equates to 128 visible patches with a patch size of 16, or 512 with a patch size of 8. We thus use these absolute values as an equivalent when masking with MultiMAE.

C. Results

Table I shows the mean AUC classification results from the 5-fold cross-validation experiment. Without pre-training, all backbones perform similarly, with AUCs from mid-60% to mid-70%, with ResNet getting the highest AUC of 0.752 when using T1 alone. BYOL significantly improves the AUC of 3D ResNet-18 on T2 to 0.838 and raising the AUC on T1 + T2 from 0.730 to 0.839. However, it does not improve the performance for T1 and multi-modal provides no increase over T2 alone. Similarly, when applied to ViT-B/8 or 16, MAE does not show improvements for T1, but for T2 and T1 + T2, significant improvements can be observed, with the best result of 0.875 AUC when using 3D ViT-B/8 on T2. There is a considerable difference of almost 9% between the best results of 3D ViT-B/16 and 3D ViT-B/8, which shows that reducing the patch size to be at most equal to the embedding size is important when configuring ViT.

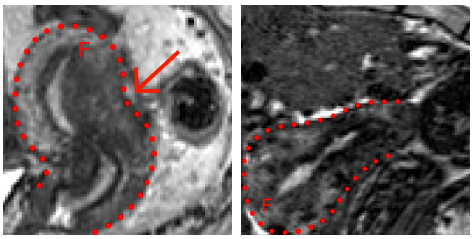

Fig. 4. Difficult positive case (Left), and difficult negative case (Right). Uterus outlined in green with the uterine fundus indicated with 'F'. Arrow points to the obliterated POD region.

Also in Table I, we can see that unlike MAE, MultiMAE performs poorly for our task. MultiMAE does not improve the AUC of 3D MultiViT/16, while the AUC of 3D MultiViT/8 increases only slightly from 0.661 to 0.693. This poor outcome may be explained by analysing the other results. It can be seen that T1 generally performs worse than T2, especially for the most accurate models, which indicates that T1 has little additional information for endometriosis classification over T2. By combining T1 and T2 with early fusion, the backbone has to extract features for classification from an input space with double the number of dimensions, without an increase of the amount of relevant information in the input. An alternate explanation is that MultiViT struggles with spatially unaligned inputs. MultiViT can distinguish between modalities through learnt biases in the linear projections, but as the positional embeddings are added after the linear projections, they cannot be ignored when combining inter-modality features while still using them when extracting intra-modality features. This adds additional complexity when learning a family of affine functions to relate the modalities' coordinate spaces, and may not work well on small datasets.

We find several cases that are exceptionally challenging to classify. Figure 4 shows examples of difficult positive and negative cases. For the positive case, only one of the models produces a correct classification. Similarly, only one model correctly classifies the negative case. For both cases, the model is the ViT-B/8 with MAE pre-training on T2.

V. Conclusion

We show that pre-training can greatly improve the performance of endometriosis classifiers. In particular, self-supervised pre-training based on MAE can increase the performance of a ViT classifier such that it surpasses ResNet on a small dataset, even though it has a weaker inductive bias. Further, we find that the relative difference between the embedding size and patch size is important. Our results indicate that the embedding size must be no smaller that the patch size. The poor performance of the evaluated methods on T1 indicates it may not be suitable for POD classification. This finding can also explain the poor performance of early fusion with MultiViT. Additional modalities may be detrimental to the performance of early fusion vision transformers if they do not contain unique information relevant to the task. These results also suggest that late fusion architectures are more robust to this phenomenon. Another explanation for the poor performance of MultiViT is that it struggles to learn a transform function to relate the coordinate spaces of unaligned modalities from a small dataset. Finally, the resulting 0.875 AUC of 3D ViT-B/8 on T2 with MAE pre-training shows the automation of endometriosis classification from MRI is possible. Further work is still required to improve the performance of these systems before they can be considered for clinical use.

References

[1] Parveen Parasar et al., "Endometriosis: epidemiology, diagnosis and clinical management," *Current obstetrics and gynecology reports*, vol. 6, no. 1, pp. 34–41, 2017.

[2] Amy L Shafrir et al., "Risk for and consequences of endometriosis: a critical epidemiologic review," *Best practice & research Clinical obstetrics & gynaecology*, vol. 51, pp. 1–15, 2018.

[3] Paolo Vercellini et al., "Endometriosis: pathogenesis and treatment," *Nature Reviews Endocrinology*, vol. 10, no. 5, pp. 261–275, 2014.

[4] Vicki Nisenblat et al., "Imaging modalities for the non-invasive diagnosis of endometriosis," *Cochrane Database of Systematic Reviews*, , no. 2, 2016.

[5] Xin Gao et al., "Economic burden of endometriosis," *Fertility and sterility*, vol. 86, no. 6, pp. 1561–1572, 2006.

[6] Kaiming He et al., "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2021.

[7] Jean-Bastien Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[8] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] Alexander Kolesnikov et al., "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1920–1929.

[10] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.

[11] Ting Chen et al., "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22243–22255, 2020.

[12] Mathilde Caron et al., "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[13] Ze Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[14] Hangbo Bao, Li Dong, and Furu Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He, "Masked autoencoders as spatiotemporal learners," *arXiv preprint arXiv:2205.09113*, 2022.

[16] Zekai Chen et al., "Masked medical image analysis," *arXiv preprint arXiv:2204.11716*, 2022.

[17] Wenzhong Guo, Jianwen Wang, and Shiping Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[19] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir, "Multimae: Multi-modal multi-task masked autoencoders," *arXiv preprint arXiv:2204.01678*, 2022.

[20] Yao-Hung Hubert Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[21] Jiawei Chen and Chiu Man Ho, "Mm-vit: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1910–1921.

[22] Jue Jiang et al., "Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit)," *arXiv preprint arXiv:2205.10342*, 2022.

[23] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna, "Self pre-training with masked autoencoders for medical image analysis," *arXiv preprint arXiv:2203.05573*, 2022.

[24] Stefano Guerriero et al., "Artificial intelligence (ai) in the detection of rectosigmoid deep endometriosis," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 261, pp. 29–33, 2021.

[25] Visalaxi Sankaravadivel et al., "Symptoms based endometriosis prediction using machine learning," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3102–3109, 2021.

[26] Hsiang-Chun Dong et al., "Using deep learning with convolutional neural network approach to identify the invasion depth of endometrial cancer in myometrium using mr images: a pilot study," *International Journal of Environmental Research and Public Health*, vol. 17, no. 16, pp. 5993, 2020.

[27] Gabriel Maicas, Mathew Leonardi, Jodie Avery, Catrina Panuccio, Gustavo Carneiro, M Louise Hull, and George Condous, "Deep learning to diagnose pouch of douglas obliteration with ultrasound sliding sign," *Reproduction and Fertility*, vol. 2, no. 4, pp. 236–243, 2021.

[28] Minmin Yang et al., "Diagnostic efficacy of ultrasound combined with magnetic resonance imaging in diagnosis of deep pelvic endometriosis under deep learning," *The Journal of Supercomputing*, vol. 77, no. 7, pp. 7598–7619, 2021.

[29] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.