# CONTINUAL PRETRAINING PROSTATE MRI VIT ENCODER FOR TASK ALIGNMENT

*Bilal A. Sidiqi*[1,2], *Yipei Wang*[1,2], *Shaheer U. Saeed*[1,2], *Shiqi Huang*[1,2], *Daniel C. Alexander*[2,3], *Yipeng Hu*[1,2]

[1]Department of Medical Physics and Biomedical Engineering,
University College London, London, UK
[2]Hawkes Institute, University College London, London, UK
[3]Department of Computer Science, University College London, London, UK

## ABSTRACT

Building foundation models for MRI has emerged as a promising direction for advancing medical image analysis and improving accessibility of task specific models. Existing approaches primarily focus on effective pretraining strategies and evaluating their performance across diverse downstream tasks. While these methods show performance gains, limited attention has been given to understanding task alignment between pretraining and downstream objectives. Unlike language models, where the two stages are naturally aligned, such alignment is less evident in vision based applications. Motivated by this gap, we explore continual pretraining. Specifically, continually pretraining a pretrained ViT encoder using Low-Rank Adaptation (LoRA) on a prostate segmentation task to enhance task alignment and downstream performance. We further perform layer-wise representational analysis using Centered Kernel Alignment (CKA) to assess representational shifts and their correlation with task outcomes. Our experiments show improved performance in 3 of 6 downstream tasks and comparable results in the others. Our Feature analysis reveals that, for segmentation tasks, continually pretrained encoder features exhibit greater similarity to those of the finetuned encoder than the original pretrained encoder, correlating with observed performance gains.

***Index Terms***— Foundation model, continual pretraining, geometric feature analysis

## 1. INTRODUCTION

Prostate MRI has an important role in reducing number of biopsies and damage to surrounding tissues during clinical diagnosis [1]. With the help of AI models, automation of information extraction, such as lesion segmentation, can further enhance clinical decision making [2]. Traditionally, independent models have been developed for different tasks, which is resource intensive, limits practical deployment, and hinders generalisation to out-of-distribution data [3]. Motivated by the success of large language models (LLMs), recent efforts to develop large scale medical vision models have shown considerable promise but continue to face several limitations. Notably, features learned through both general-purpose [4] and domain-specific [5] pretraining have not consistently translated to strong performance in medical downstream tasks, often requiring extensive task-specific fine-tuning. Although various pretraining strategies such as MAE, DINO, and CLIP have been explored [6, 7], each offering task-dependent advantages, they still seem to lack balanced representations for capturing the shared semantics across downstream tasks [8].

In this work, we propose a continual pretraining approach aimed at improving the learning of shared downstream task representations. Continual pretraining generally involves further pretraining an already pretrained encoder to better align its learned features with those relevant to downstream tasks. A recent study [8] explored a similar idea by continually pretraining self-supervised DINOv2 and supervised CLIP models on domain-specific data at scales of 30k, 50k, 100k, 1M, and 1.3M samples, reporting consistent performance improvements with larger in-domain datasets—where, in some cases, as few as 30k samples outperformed the original pretrained weights. While our dataset is considerably smaller, we nevertheless observe improved performance in several in-domain and out-of-domain downstream tasks.

Specifically, we employ a pretrained Vision Transformer (ViT) encoder [9] trained using 5,524 3D prostate MRI scans on a self-supervised Masked Autoencoder (MAE) objective to capture multi-level general features from MRI data. We continually pretrain this encoder using Low-Rank Adaptation (LoRA) [10] on a prostate segmentation task comprising 396 training and 113 test samples. The prostate segmentation task was chosen due to its close relevance to several downstream tasks, including prostate lesion and anatomical segmentation, PI-RADS score classification, and prostate volume regression. Across six downstream task examples, we observe performance improvements in three tasks following continual pretraining. Two of which involve segmentation, highlighting the significance of task similarity for achieving better alignment during continual pretraining.

To further analyse representational changes, we utilise the Centered Kernel Alignment (CKA) similarity metric [11] to quantify pairwise feature representation similarities across

three encoders: the pretrained baseline, the fine-tuned model (initialised from the pretrained encoder), and our continual pretrained encoder. We analyse the difference between CKA(pretrained, fine-tuned) and CKA(continual-pretrained, fine-tuned) to assess whether continual pretraining shifts the feature geometry closer to that of the fine-tuned model, indicating improved task alignment. Furthermore, we examine whether the change in representational similarity correlates with the downstream performance difference between fine-tuned models initialised from pretrained and continually pretrained encoders. Our results show that this correlation holds when a measurable representational similarity shift is observed, suggesting that geometric feature alignment is indeed important in improved downstream performance.

## 2. METHOD

### 2.1. Applying LoRA to a Pretrained Encoder

We propose a continual pretraining framework in which a 3D Vision Transformer (ViT) encoder, pretrained on a self-supervised Masked Autoencoder (MAE) objective, is further aligned to downstream tasks using Low-Rank Adaptation (LoRA) through supervised prostate segmentation. Specifically, LoRA is integrated into the attention mechanism of each transformer block within the encoder. This modifies the original query, key, value, and output projection matrices $(W_Q, W_K, W_V, W_O)$ by introducing trainable low-rank update matrices, formulated as:

$$W = W_0 + \alpha BA, \tag{1}$$

where $W_{0\in\{Q,K,V,O\}}$ represents the original weight matrix, $A$ and $B$ are low-rank matrices capturing essential modes of variation within the weight space, and $\alpha$ is a scaling factor controlling the adaptation magnitude. The adapted transformer block can be expressed as:

$$\hat{Z}^{(\ell)} = \text{LayerNormalisation}\big(Z^{(\ell-1)}\big), \tag{2}$$

$$Q = \hat{Z}^{(\ell)}W_Q, \quad K = \hat{Z}^{(\ell)}W_K, \quad V = \hat{Z}^{(\ell)}W_V, \tag{3}$$

$$\text{Attention}(Q,K,V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_h}}\Big)V, \tag{4}$$

$$Z^{(\ell,1)} = Z^{(\ell-1)} + \text{Attention}(Q,K,V)W_O, \tag{5}$$

$$Z^{(\ell)} = Z^{(\ell,1)} + \text{MLP}\Big(\text{LayerNormalisation}\big(Z^{(\ell,1)}\big)\Big), \tag{6}$$

where $\hat{Z}^{(\ell)}$ denotes the normalized layer output, $d_h$ is the dimensionality of the query/key vectors, $W_{0\in\{Q,K,V,O\}}$ are the LoRA-adapted weight matrices and MLP is Multilayer Perceptron.

### 2.2. Continual Pretraining

During continual pretraining, the pretrained encoder weights $W$ are frozen, and an adapter module is used to connect a Unified Perceptual Parsing (UperNet) decoder for segmentation. The adapter receives hidden states $\{H_i\}_{i=1}^4$ from selected layers of the ViT encoder and reshapes each token sequence into a 3D feature map $X_i = \text{Reshape}(H_i) \in R^{B\times C\times D\times H\times W}$. The UperNet decoder integrates these multiscale feature maps by downsampling the first two and upsampling the last to align them to the spatial scale of the third map. The fused representation is then passed through a softmax classifier to produce voxel-level segmentation. LoRA parameters are trained to adapt the encoder from generating general features learned during pretraining to producing task-aligned representations more suitable for related downstream tasks.

### 2.3. Downstream Tasks

Following continual pretraining, the encoder's original weights are unfrozen and evaluated across a range of downstream tasks using both in-domain (identical to continual pretraining data) and out-of-domain datasets, as summarised in Table 1. To isolate encoder effects, we employ simple task specific heads. A lightweight two layer linear head is attached for prostate zone classification and prostate volume regression. For lesion and prostate zone segmentation, we employ the same UperNet decoder architecture. Finally, a super-resolution module from [12] is attached to the encoder to enhance low-resolution $T_2$ inputs by a factor of four.

### 2.4. Geometric Feature Analysis Metric

To analyse layer-wise representational similarity across the pretrained, continual-pretrained, and fine-tuned encoders, we employ the Centered Kernel Alignment (CKA) similarity metric. CKA quantifies the degree to which two sets of feature representations encode similar geometric structures in their respective feature spaces. We compute pairwise CKA values between encoder layers (e.g. $l_1$ and $l_2$) for tasks exhibiting a performance difference of approximately $0.2\%$ or greater. For given layer outputs $Z_1^{(l_1)}$ and $Z_2^{(l_2)}$ from encoders $Z_1$ and $Z_2$, the linear CKA similarity is defined as:

$$\text{CKA}\big(Z_1^{(l_1)}, Z_2^{(l_2)}\big) = \frac{\|{Z_1^{(l_1)}}^\top Z_2^{(l_2)}\|_F^2}{\|{Z_1^{(l_1)}}^\top Z_1^{(l_1)}\|_F \, \|{Z_2^{(l_2)}}^\top Z_2^{(l_2)}\|_F}, \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In our experiments, CKA similarities are computed for layers 3, 5, 7, and 11 across the three encoders to evaluate how fine-tuning and continual pretraining influence internal feature representations relative to the pretrained baseline.

## 3. EXPERIMENTS

### 3.1. Dataset and Experimental Setup

Four datasets were used in this study: PROMIS for continual pretraining, Anatomy for downstream anatomy segmentation, RISK for PIRADS score classification, and UCL Set

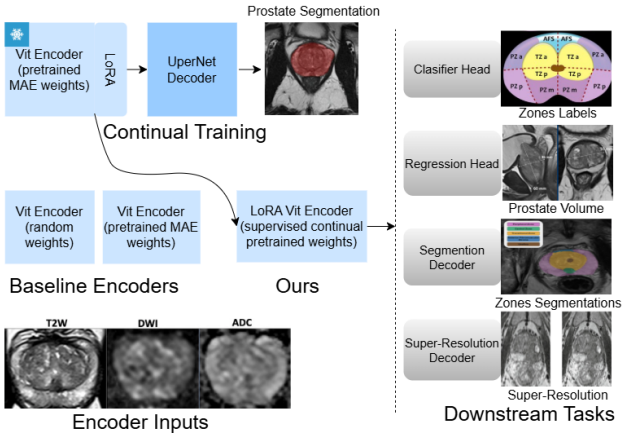


**Fig. 1**. Overview of the proposed continual pretraining framework, integrating LoRA within a pretrained ViT encoder to continually pretrain the encoder for better task alignment and downstream performance.

for Prostate volume regression [13, 14, 15]. Their respective data splits, input modalities, and target outputs are summarised in Table 1. All models were trained with an input size of $(64, 256, 256)$, except for the super-resolution task where the depth dimension $(64)$ was reduced by a factor of four. The inputs underwent standard preprocessing and augmentation procedures, including normalisation, centre cropping, random rotation, flipping, zooming, and intensity variation.

All experiments were conducted on an NVIDIA A40 GPU for 100 epochs. The Adam optimiser was used across all tasks, with Dice loss applied to segmentation, cross-entropy loss to classification, mean squared error (MSE) to regression, and mean absolute error (MAE) to super-resolution. Batch sizes were set to $4$ for segmentation and super-resolution tasks, and $16$ for classification and regression tasks.

| Dataset | Data Split | MRI Input | Output |
|---|---|---|---|
| PROMIS | (396,57,113) | T2W, DWI, ADC | Lesion Mask |
| Anatomy | (412,60,117) | T2W | Prostate Zones Mask |
| RISK | (718,103,207) | T2W, DWI, ADC | PIRADS Score (2–5) |
| UCL Set | (1386,178,387) | T2W | Prostate Volume |

**Table 1**. Overview of datasets, data splits, and corresponding input–output configurations.

### 3.2. Results

Model performance was compared across three configurations sharing identical architectures but differing in weight initialisation. The first, *ViT Random Weights*, was trained from scratch. The second, *ProFound ViT OMW*, initialised with pretrained encoder weights, served as the baseline. The third, *ProFound ViT Ours*, utilised our continually pretrained encoder weights.

**Performance on Classification Tasks:** Two classification tasks were evaluated. The first employed the RISK dataset for four-class PIRADS score classification (2–5), where higher scores indicate increased malignancy risk. The second used the PROMIS dataset for 20-zone prostate classification. As shown in Table 2, for PIRADS classification, the baseline model achieved the highest overall AUC, outperforming our model by approximately $0.4\%$. However, for AUC thresholds $\geq 3$ and $\geq 4$, our model surpassed the baseline by over $5\%$ in AUC$\geq 3$, suggesting improved discrimination of higher-risk PIRADS scores. For zone classification, the baseline outperformed our model by $2.3\%$ on average.

Feature representation analysis (Section 3.2.5) revealed that encoder representations from the continually pretrained and finetuned models were highly similar to the pretrained encoder. This suggests that for classification tasks, globally learned representations during pretraining remain largely transferable, and performance differences are likely influenced by subtle variations in discriminative feature directions [16].

**Performance on Regression Task:** As presented in Table 3, regression results showed a marginal $0.2\%$ decrease compared to the baseline, indicating limited benefit of continual pretraining for this specific alignment task.

**Performance on Segmentation Tasks:** We evaluated two segmentation tasks: lesion segmentation using the PROMIS dataset and prostate zone segmentation using the Anatomy dataset (eight zones in total). Table 4 shows that our model outperformed the baseline by approximately $0.5\%$ in average Dice score for lesion segmentation and $0.1\%$ for zone segmentation. These improvements suggest that continual pretraining on a similar segmentation task enhances alignment for downstream segmentation objectives. This observation is further examined in Section 3.1.1, where representational similarity is analysed.

**Performance on Super-Resolution Task:** Table 5 presents results for the super-resolution task, where the baseline model achieved a slightly higher SSIM ($0.1\%$ improvement). Similar to the regression case, continual pretraining did not yield improved task alignment for this setting.

| | AUC | AUC $\geq$3 | AUC $\geq$4 | Avg Std |
|---|---|---|---|---|
| **PIRADS Classification (RISK Data)** | | | | |
| ViT Random Weights | 58.81 | 69.98 | 62.82 | 2.51 |
| ProFound ViT OMW | **61.36** | 74.76 | **64.03** | 2.56 |
| ProFound ViT Ours | 60.94 | **79.92** | 64.00 | 2.63 |
| **Zone Classification (PROMIS Data)** | | | | |
| ViT Random Weights | 64.41 | – | – | 1.97 |
| ProFound ViT OMW | **72.64** | – | – | 1.75 |
| ProFound ViT Ours | 70.36 | – | – | 1.86 |

**Table 2**. Results for classification tasks.

| **Prostate Volume Regression (UCL Data)** | | |
|---|---|---|
| | MSE | Std |
| ViT Random Weights | 0.038 | 0.028 |
| ProFound ViT OMW | **0.013** | 0.009 |
| ProFound ViT Ours | 0.015 | 0.012 |

**Table 3**. Results for regression task.

| **Lesion Segmentation (PROMIS Data)** | | | |
|---|---|---|---|
| | **Avg Dice** | **Class Dice** | **Std** |
| ViT Random Weights | 0.216 | – | 0.186 |
| ProFound ViT OMW | 0.248 | – | 0.203 |
| ProFound ViT Ours | **0.253** | – | 0.206 |
| **Zone Segmentation (Anatomy Data)** | | | |
| | **Avg Dice** | **Class Dice** | **Std** |
| ViT Random Weights | 0.830 | 0.910, **0.913**, **0.868**, 0.771, 0.875, 0.907, 0.750, 0.646 | 0.093 |
| ProFound ViT OMW | 0.835 | 0.916, 0.908, 0.866, **0.782**, 0.880, 0.912, 0.765, 0.653 | 0.095 |
| ProFound ViT Ours | **0.836** | **0.917**, 0.909, 0.866, 0.780, **0.881**, **0.912**, **0.765**, **0.657** | 0.095 |

**Table 4**. Results for segmentation tasks.

| **T2 MRI Super-Resolution (PROMIS Data)** | | |
|---|---|---|
| | SSIM | Std |
| ViT Random Weights | 0.900 | 0.009 |
| ProFound ViT OMW | **0.931** | 0.005 |
| ProFound ViT Ours | 0.930 | 0.005 |

**Table 5**. Results for super-resolution task.

**Feature Representation Analysis:** Table 6 presents the layer wise and average CKA similarity scores between encoder pairs across four selected layers (3, 5, 7, and 11). For the pretrained–continual pair, the encoders exhibit high similarity across all layers, with a gradual decrease in deeper layers, suggesting that LoRA primarily refines high-level semantic representations while inducing task-specific adaptations in later layers.

In the lesion segmentation task, finetuned encoder representations (using pretrained weights) show decreasing similarity to both pretrained and continual pretrained encoders with layer depth. Notably, at the deepest layer, similarity with the continual pretrained encoder increases by approximately $0.4\%$, aligning with the observed improvement in segmentation performance (Table 4). This supports the hypothesis that continual pretraining facilitates better task-aligned representation learning.

For the PIRADS classification task, the finetuned encoder remains nearly identical to both pretrained and continual pretrained encoders, implying minimal representational shifts during finetuning. This suggests that classification primarily benefits from globally stable representations rather than layer specific adaptations.

| **Encoder Pair** ($Z_1 - Z_2$) | L3 | L5 | L7 | L11 | **Avg.** |
|---|---|---|---|---|---|
| Pretrained–Continual | 0.997 | 0.996 | 0.997 | 0.995 | 0.996 |
| **CKA Similarity for Lesion Segmentation** | | | | | |
| Pretrained–Finetuned | **0.918** | **0.915** | **0.893** | 0.340 | **0.767** |
| Continual–Finetuned | 0.915 | 0.910 | 0.890 | **0.344** | 0.765 |
| **CKA Similarity for PIRADS Classification** | | | | | |
| Pretrained–Finetuned | **1.000** | **1.000** | **1.000** | **0.993** | **0.998** |
| Continual–Finetuned | 0.997 | 0.997 | 0.997 | 0.989 | 0.995 |

**Table 6**. Layer-wise and average CKA similarity between encoder pairs. "Pretrained" denotes the encoder with pretrained weights; "Continual" indicates the pretrained encoder further trained with LoRA on the prostate segmentation task; "Finetuned" refers to the encoder finetuned on downstream tasks (lesion segmentation or PIRADS classification).

## 4. CONCLUSION

In this work, we investigated a novel direction of continual pretraining as a means of advancing foundation model development for prostate MRI analysis. The objective was to enhance task alignment and improve downstream task performance. Our experimental results demonstrate improved and comparable performances in all 6 downstream tasks, two of which involve segmentation, suggesting that task similarity during continual pretraining contributes to improved alignment. Our Feature analysis results show that, for segmentation tasks, the output feature representations of the continual pretrained encoder exhibit greater similarity to those of the finetuned (initialised from the pretrained) encoder compared to the pretrained encoder itself. This increased representational closeness correlates with the observed performance improvements when finetuning from continual pretrained weights. Conversely, for classification tasks, the representational geometry exhibited minimal change during finetuning, likely due to a stronger focus on fine grained discriminative details, therefore limiting the ability to infer a clear correlation with performance.

Future work may explore integrating a Mixture-of-Experts (MoE) approach in the later layers of the encoder to facilitate multi-task continual pretraining and achieve more generalisable task alignment. Moreover, employing alternative representational similarity metrics could provide deeper insights into the impact of continual pretraining on feature representations.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Ethical approval was granted by the Data Trust Committee, University College London Hospitals (24-04-2024/IRAS ID 299136).

## 6. REFERENCES


[1] Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent automatic segmentation algorithms of mri prostate regions: A review," *IEEE Access*, vol. 9, pp. 97 878–97 905, 2021.

[2] B. Turkbey and M. A. Haider, "Deep learning-based artificial intelligence applications in prostate mri: brief summary," *The British Journal of Radiology*, vol. 95, no. 1131, p. 20210563, 2022.

[3] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen *et al.*, "Sam-med3d: towards general-purpose segmentation models for volumetric medical images," in *European Conference on Computer Vision*. Springer, 2024, pp. 51–67.

[4] J. Wu, R. Fu, H. Fang, Y. Liu, Z.-Y. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *Medical image analysis*, vol. 102, p. 103547, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258309597

[5] E. Burshtein, N. Cahan, L. Ayzenberg, and H. Greenspan, "Cxr-dino: paving the way for a medical vision foundation model through self-supervised learning in chest x-ray analysis," in *Medical Imaging*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277756367

[6] G. Campanella, R. Kwan, E. Fluder, J. Zeng, A. Stock, B. Veremis, A. D. Polydorides, C. Hedvat, A. Schoenfeld, C. Vanderbilt *et al.*, "Computational pathology at health system scale–self-supervised foundation models from three billion images," *arXiv preprint arXiv:2310.07033*, 2023.

[7] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L. Teng, D. Liu, X. Li, Z. Cui, Q. Wang, and D. Shen, "Clip in medical imaging: A comprehensive survey," *ArXiv*, vol. abs/2312.07353, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266174283

[8] M. Ilse, H. Sharma, A. Schwaighofer, S. Bond-Taylor, F. Pérez-García, O. Melnichenko, A.-M. G. Sykes, K. K. Horst, A. Khandelwal, M. Reynolds *et al.*, "Data scaling laws for radiology foundation models," *arXiv preprint arXiv:2509.12818*, 2025.

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

[10] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *ArXiv*, vol. abs/2106.09685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235458009

[11] S. Kornblith, M. Norouzi, H. Lee, and G. E. Hinton, "Similarity of neural network representations revisited," *ArXiv*, vol. abs/1905.00414, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:141460329

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 295–307, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6593498

[13] T. Marsden, N. McCartan, L. Brown, M. Rodriguez-Justo, T. Syer, G. Brembilla, M. Van Hemelrijck, T. Coolen, G. Attard, S. Punwani *et al.*, "The reimagine prostate cancer risk study protocol: A prospective cohort study in men with a suspicion of prostate cancer who are referred onto an mri-based diagnostic pathway with donation of tissue, blood and urine for biomarker analyses." *Plos one*, vol. 17, no. 2, p. e0259672, 2022.

[14] Y. Li, Y. Fu, I. J. Gayo, Q. Yang, Z. Min, S. U. Saeed, W. Yan, Y. Wang, J. A. Noble, M. Emberton, M. J. Clarkson, H. Huisman, D. C. Barratt, V. A. Prisacariu, and Y. Hu, "Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration," *Medical Image Analysis*, vol. 90, p. 102935, 2023.

[15] H. U. Ahmed, A. El-Shater Bosaily, L. C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R. G. Hindley, A. Freeman, A. P. Kirkham, R. Oldroyd, C. Parker, and M. Emberton, "Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study," *The Lancet*, vol. 389, no. 10071, pp. 815–822, 2017.

[16] Y. Ruiping, L. Kun, X. Shaohua, Y. Jian, and Z. Zhen, "Vit-upernet: a hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation," *Complex Intelligent Systems*, vol. 10, 02 2024.