Imperial College London

FINAL REPORT FOR MENG PROJECT

IMPERIAL COLLEGE LONDON

DEPARTMENT OF BIOENGINEERING

Self-Supervised Learning of Swin Transformer for Automated Cartilage Segmentation on 7 Tesla Knee MRI

Tin Lam Cheng
CID: 01717301

Supervisor(s):
Dr. Neal K. Bangerter

Submitted in partial fulfilment of the requirements for the award of MEng in Biomedical Engineering from Imperial College London

Word Count: 5815

Date: June 14, 2023

Abstract

Knee osteoarthritis (KOA) is a common degenerative joint disease associated with age, resulting from the progressive loss of cartilage. This study focuses on leveraging the ultra-high resolution 7 Tesla (7T) magnetic resonance imaging (MRI) of the knee, to develop an automated cartilage segmentation system to facilitate cartilage thickness quantification for early-stage KOA detection. While the limited availability of annotated 7T MRI hinders traditional supervised deep learning performance, self-supervised learning (SSL) enables model extraction and learning of image features from unannotated data, reducing the reliance on costly manual annotations.

We utilized the 2D Swin-UNETR model, combining a Swin Transformer (SwinT) and U-Net for 7T MRI cartilage segmentation. The SwinT was first pretrained on a large amount of unlabelled low-resolution 3T MRI with SSL, and subsequently fine-tuned on the small labelled 7T MRI dataset for segmentation. SSL pretext tasks including image inpainting, image rotation prediction, and contrastive learning were applied, revealing that the former two tasks provided higher effectiveness in facilitating downstream segmentation. Results showed a slight improvement by the SSL-pretrained model with multiple pretext tasks applied (Dice Similarity Coefficient (DSC) of 0.938), compared to the purely supervised model (0.928). It also showed data efficiency by reducing annotation effort by approximately 40%. Model overfitting onto the small and low-variability 7T MRI dataset remains a challenge for future investigation. This SwinT-based model can also be extended to 3D whole-knee MRI segmentation for more precise cartilage quantification in the future.

Acknowledgements

I would like to express my heartfelt appreciation to my supervisor, Dr. Neal K. Bangerter, for his invaluable guidance and continuous support throughout my research journey. His advice and expertise have been instrumental in assisting me to complete my research.

I am grateful to Krithika Balaji and Michael Mendoza for their assistance in setting up the computing environment and facilitating access to the necessary datasets.

I would also like to extend my special thanks to Joonsu Gha for generously sharing his professional insights and experiences in shaping my research direction.

Lastly, I would like to acknowledge the support of the staff from the Research Computing Service team at Imperial College London for their assistance in setting up the HPC environment for GPU access.

Contents

1 Introduction

Cartilage is a smooth tissue coating two ends of bones at joints, serving as a shock absorber to reduce the impact of forces during body movement. Aging leads to the gradual wearing of cartilage, resulting in joint inflammation and the development of osteoarthritis (OA) [1]. OA affects approximately 7% of the global population (500 million) [2], with knee osteoarthritis (KOA) being one of the most common types. The initial KOA diagnostic procedure involves radiographic examinations of the knee with the severity classified according to the Kellgren-Lawrence (KL) grading, which is a measure based on the narrowing of the joint space [3]. However, this approach often detects KOA at a later stage when the joint space narrowing is already pronounced and symptoms have developed. At this point, the damage to the joint is irreversible. Current treatments primarily focus on managing pain, controlling the severity, or resorting to joint replacement surgery in severe cases. There is a critical need to investigate early-stage KOA detection with imaging techniques.

While conventional radiographs provide a limited anatomical view for KOA [4], there is an increasing amount of research dedicated to KOA diagnoses using Magnetic Resonance Imaging (MRI), due to its capability in visualizing 3D knee joint structures and cartilage lesions [5]. In particular, the establishment of 7 Tesla (7T) ultra-high resolution MRI technology in 2017 provides a much higher spatial and temporal resolution for more reliable visualisation of structures and tissues [6]. With cartilage loss being a biomarker for KOA [7], the cartilage region of interest (ROI) can be extracted from MRI with image segmentation techniques for quantifying its volume and thickness, allowing an accurate identification of the amount of cartilage loss for early-stage KOA diagnosis.

Emerging deep learning (DL) technologies over the past years have gained remarkable success in image segmentation [8], mainly with convolutional neural networks (CNN). CNNs employ a hierarchical approach to extract latent features, ranging from fine-level edges to intricate objects within the image [9]. However, most successful CNN-based methods rely on fully supervised training, which requires a substantial amount of annotated data. This poses a challenge in the medical domain, where it is difficult to acquire large quantities of medical images, along with their annotations that require domain expertise. The gold standard segmentation labels are only achieved manually by radiologists, which is both expensive and time-consuming [10]. Particularly for 7T ultra-high resolution MRI, precisely delineating boundaries and ROI becomes notably more laborious due to the increased volume and level of structural details. As a result, there is a scarcity of annotated 7T MRI available for effective supervised DL.

Existing solutions for addressing the small-dataset segmentation problem in the medical domain primarily rely on transfer learning (TL). The model is initially pretrained on a relevant but different large dataset in a supervised manner, with their learned knowledge transferring onto the segmentation of the small target dataset by fine-tuning the model training parameters [11]. This approach is effective when both pretraining and target datasets are annotated with the same image domain and modality, as the similarity in image format and tissue structures maximizes the benefit of TL. However, curating annotated datasets for all image modalities, imaging tasks and outcomes for supervised learning is cost-prohibitive and time-consuming [12]. Therefore, it is important to develop an auto-

mated segmentation pipeline that minimizes the reliance on annotated datasets.

Self-supervised learning (SSL), combining the idea of TL, is a promising solution to the lack of large annotated medical datasets. It enables pretraining on unannotated data to extract useful feature representations that can be transferred to improve downstream segmentation performance. SSL has shown success in Natural Language Processing (NLP), especially with the use of self-attention-based Transformer models [13, 14], capable of achieving state-of-the-art results from fine-tuning models pretrained from enormous unlabelled scraped text data. There is an increasing popularity in translating similar techniques to computer vision, with the use of Vision Transformers (ViT) [15, 16]. In particular, recent research shows that ViT outperformed CNN in SSL, especially with large unannotated images with computation quadruply reduced [17, 18], showing huge potential in utilizing SSL with ViT for medical imaging applications.

In this study, we developed an automated SSL cartilage segmentation system for a scarce amount of annotated 7T knee MRI with a ViT-based model. The aim is to achieve the gold standard segmentations that facilitate future use in cartilage thickness quantification for early-stage KOA detection. Another key objective is to achieve data efficiency by maintaining model performance under a reduced amount of 7T MRI. We leveraged SSL by first pretraining on relatively large unannotated low-resolution 3T knee MRI available from the Osteoarthritis Initiative (OAI) [19], followed by supervised fine-tuning on limited 7T MRI for cartilage segmentation. Results showed highly accurate segmentations from the purely supervised model, with a slight improvement with SSL implementation, resulting in almost outlier-free segmentations. This modest improvement from SSL also underlined the need for future investigation into potential model overfitting issues.

2 Methods

2.1 Theory

2.1.1 Self-Supervised Learning (SSL)

SSL is considered an effective solution for the scarcity of annotated medical images in DL training, by enabling the network to extract and learn robust latent representations from unannotated images [20]. The SSL workflow comprises two stages, pretext and downstream. The pretext stage entails transforming an unsupervised problem into a supervised problem by generating labels from the image data itself and creating informative auxiliary pretext tasks for the model to learn about the underlying image structure and features. The acquired knowledge is subsequently transferred to the downstream target task, in specific, semantic segmentation. The model leverages the rich information captured during pretext learning to improve the accuracy and performance of the segmentation, even in scenarios with limited annotated data.

Pretext tasks can be further categorized into predictive, generative and contrastive tasks [20]. Predictive approaches treat the pretext task as a classification problem through assigning each unlabelled image with a pseudo label generated from the data itself. Common tasks explored in the medical domain include image rotation prediction [21] and patch relative position prediction [22] etc. Generative tasks involve model learning to regenerate

the same input data, such as image inpainting [23], allowing learning of representations through completing the cropped or masked pixels by observing surrounding pixel patterns. Contrastive learning [24] learns by differentiating between similar (positive) and diverse (negative) image pairs by pulling the former together and the latter apart in the latent space [25]. Further descriptions of different pretext tasks and their implementations in medical imaging are discussed in the planning report [26]. In this study, image inpainting (generative), image rotation prediction (predictive) and contrastive learning tasks were applied to investigate the model's learning effectiveness with different pretext task types to facilitate downstream target segmentation.

2.1.2 U-Net for Medical Image Segmentation

U-Net architecture has gained significant popularity in biomedical image segmentation, offering efficient and precise segmentation despite limited training data [27]. Figure 1 shows the original 2D CNN-based U-Net structure, featuring an encoder-decoder design with contracting and expanding paths interconnected by skip connections. The encoder extracts image features through downsampling, while the decoder restores image details and spatial dimensions via upsampling. Concatenated skip connections from the encoder to the decoder allows information retention from previous layers. Extensive research efforts have been devoted to the development of various U-Net variants [28] to meet the growing performance demand in medical image segmentation, with details discussed in the planning report [26].

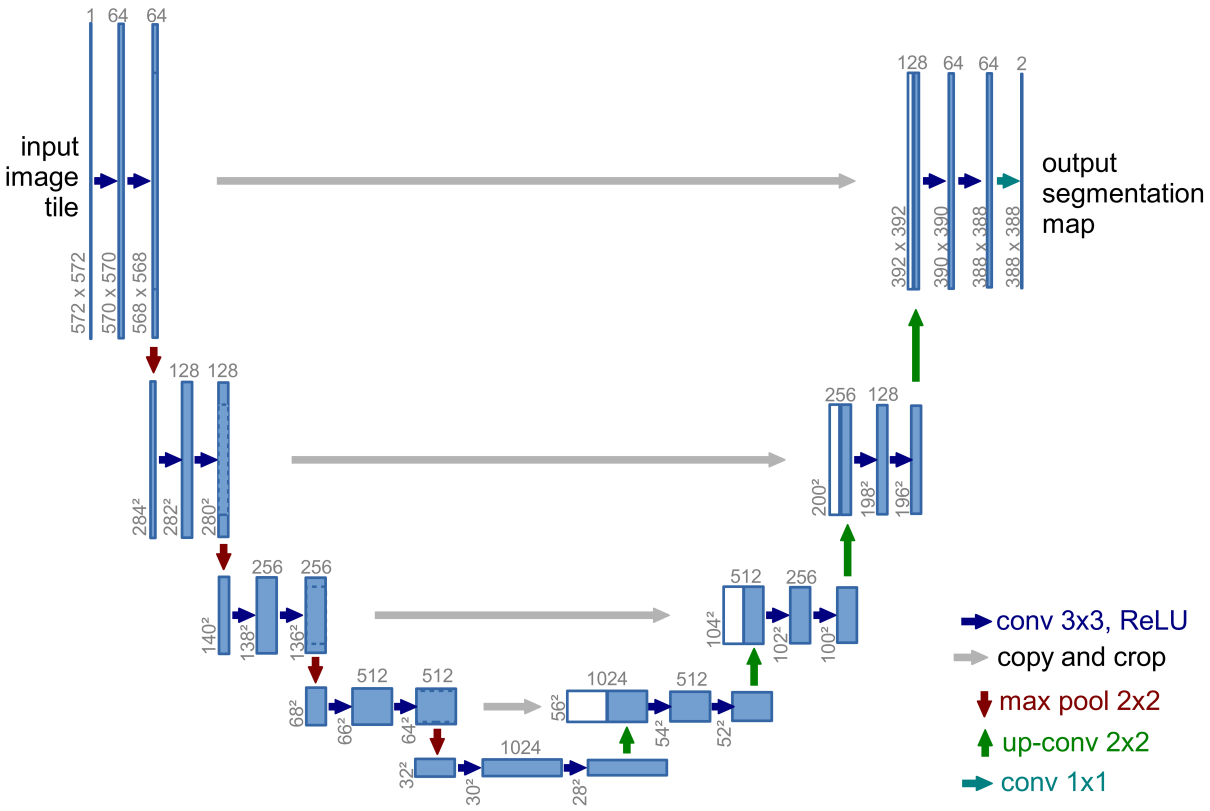


Figure 1: Structure of the 2D CNN-based U-Net [27]

2.1.3 Vision Transformer (ViT)-Based Models

Inspired by the successful applications of Transformer models in NLP [29], ViT [16] processes images in a similar way using a self-attention mechanism. As shown in Figure 2(a), images are split into patches, with each patch being treated as a 'token'. The sequence of patches is then processed through stacked Transformer encoders (Figure 2(b)). With Multi-head Self-Attention (MSA), it enables self-alignment learning to determine the relative importance of each image patch compared to the others. This patch-based learning approach also facilitates the extraction of information even without labelled data. ViT offers advantages over CNN, especially in dense prediction tasks like segmentation with large images and small ROI [30], while maintaining lower computational complexity [18].

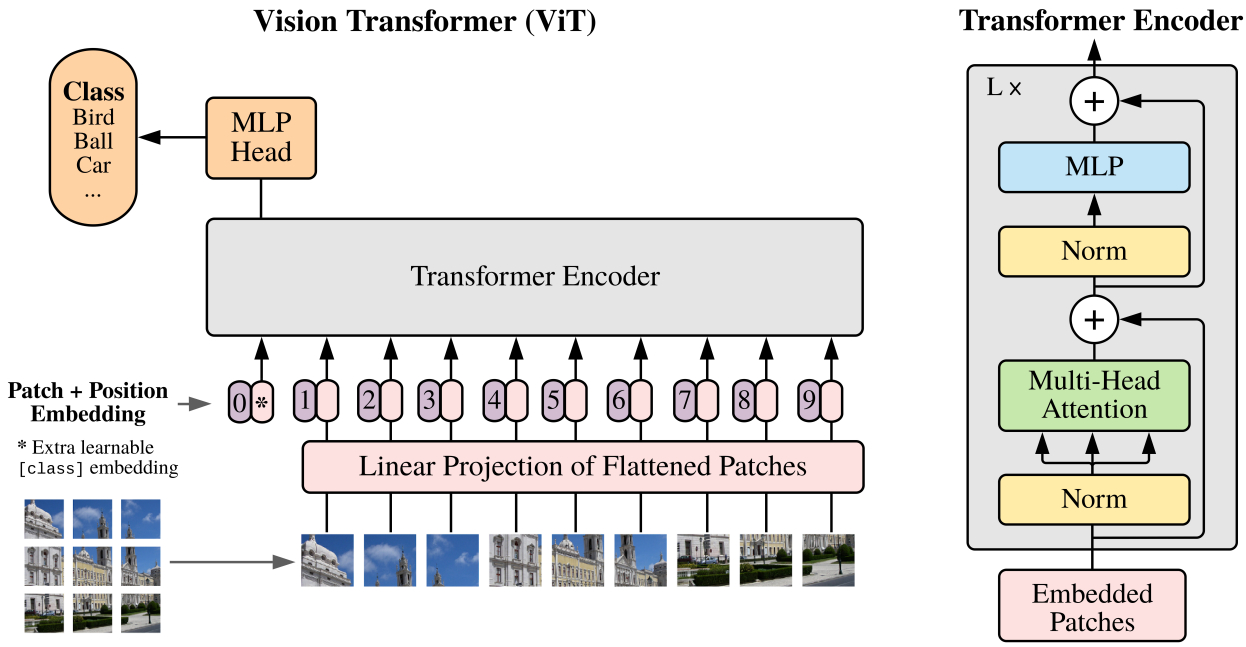


Figure 2: (a) Structure of ViT: an image is split into fixed-size patches, each of them being linearly embedded. The resulting sequence of the vectors is fed into the Transformer encoder. Multi-layer perception (MLP head) is attached for performing the designated task such as classification; (b) Structure of a Transformer encoder block: the multi-head attention (MHA) layer learns local and global dependencies of an image. The MLP encodes the outputs from MHA for the designated task [16].

Several ViT-based U-Net architectures have been developed specifically for medical image segmentation, as described in the planning report [26]. One notable ViT variant is the Swin Transformer (SwinT) [31], shown in Figure 4, which tackles the problem of large variations in visual entities while also improving the efficiency on high-resolution images. In each SwinT block, as illustrated in Figure 3, instead of the standard MSA, the Window-based MSA (W-MSA) computes attentions solely within each window to learn neighboring pixel relationships. This is followed by a Shifted-Window MSA (SW-MSA) to shift the window along the image to introduce important cross-window connections to enhance feature learning. SwinT's hierarchical architecture also allows its adaptation as the U-Net encoder, forming the Swin-UNETR proposed by Hatamizadeh et al. [32]. This architecture has demonstrated top performance on 3D computed tomography (CT) and MRI segmentations with SSL [33, 34]. Therefore, in this study, the Swin-UNETR structure was used for cartilage segmentation.

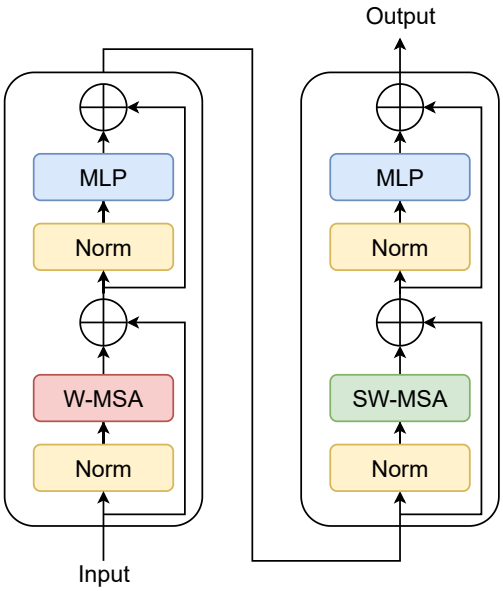


Figure 3: Structure of a Swin Transformer block [31]: It consists of two subunits. Each subunit consists of a normalization layer (Norm) and an attention module, followed by another normalization layer and MLP. The first sub-unit applies a Window-based MSA (W-MSA) module while the second sub-unit applies a Shifted Window MSA (SW-MSA) module.

2.2 The Proposed Method - SSL-Pretrained SwinT-based MRI Cartilage Segmentation System

The proposed method for single-class cartilage segmentation made use of both SwinT [31] and U-Net [27] structures. The process was divided into two stages, SSL pretraining and supervised downstream segmentation. To avoid high computational costs in 3D neural network training, instead, we employed 2D training by slicing the 3D MRIs.

2.2.1 Stage 1: SSL Pretraining

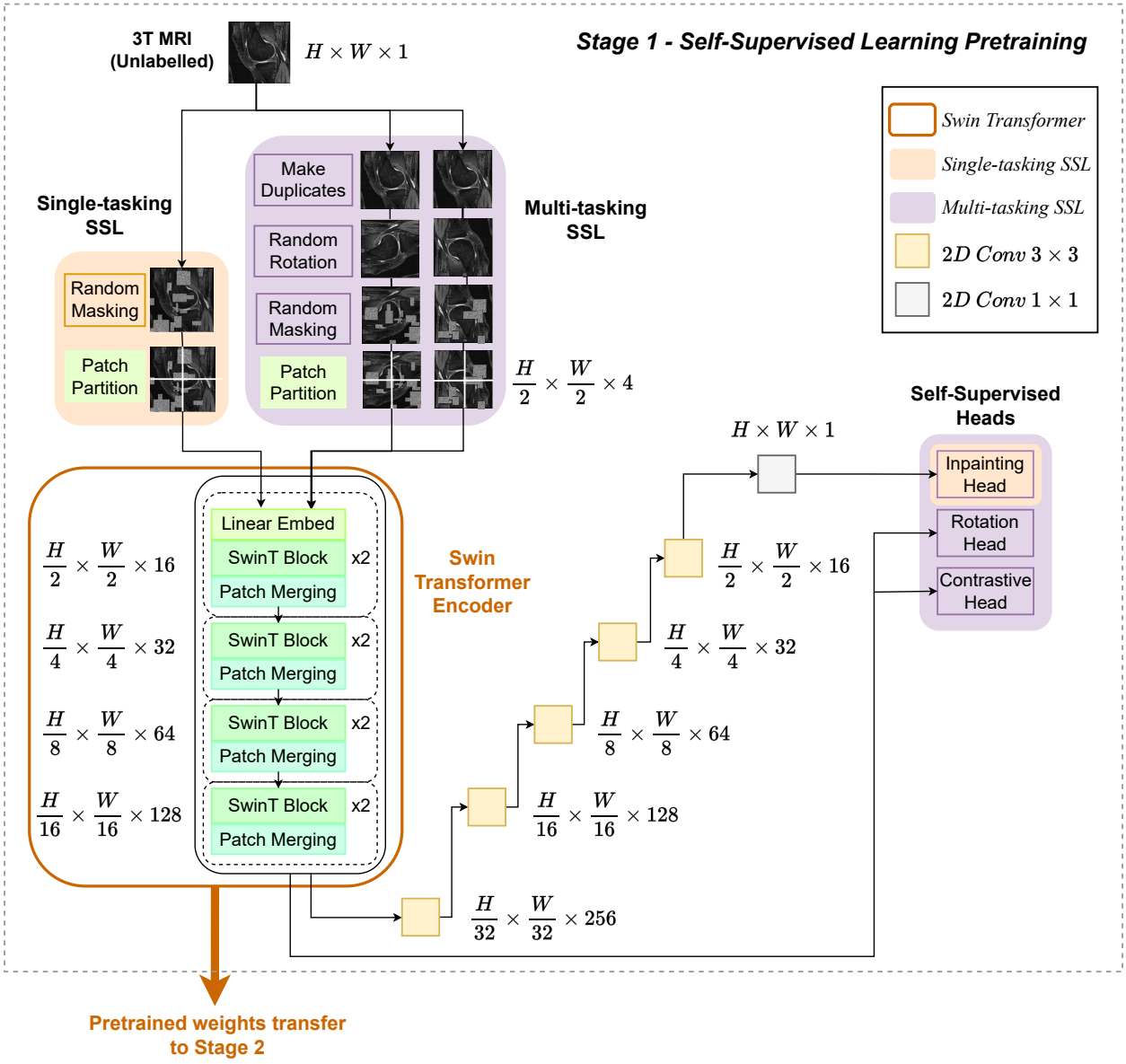


Figure 4: Illustration of the process and network architecture used in Stage 1 - SSL pretraining.

As described in Figure 4, Stage 1 involves pretraining the SwinT encoder with SSL using unannotated 3T MRI from OAI. In particular, We explored several pretext tasks for the encoder to learn image feature representations using two SSL approaches, single-tasking and multi-tasking SSL.

(i) Single-Tasking SSL: Initially, for simplicity and assessing the efficacy of SSL in enhancing downstream segmentation, we assigned only one pretext task (single-tasking), image inpainting, to the SwinT encoder. As suggested in the SSL SwinT paper [34], random patches, ranging from 0% to 30% of the image, were randomly masked at various positions for each input image. Each image was then partitioned into patches using a patch size of 2×2 with a feature dimension of $2 \times 2 \times 1 = 4$ on single channelled 3T MRI. The input image became a sequence of 4 patch tokens with a size $\left(\frac{H}{2} \times \frac{W}{2}\right)$, which were then

fed into the SwinT encoder.

Input patches were later projected onto a 16-dimensional embedding space. They were subsequently passed into a group of 2 SwinT blocks, dividing into non-overlapping windows for self-attention computations using W-MSA and SW-MSA as illustrated in Figure 3. Outputs of each SwinT group underwent patch merging, where neighboring patches were grouped and depth-wise concatenated. This effectively downsampled the input resolution by a factor of 2. This process was repeated until reaching the fourth group of SwinT blocks with a resolution of $(\frac{H}{16} \times \frac{W}{16})$. After pretraining, weights of this hierarchical encoder were saved for later use in downstream segmentation. The predicted image was successively reconstructed through upsampling with a simple CNN-based decoder, guided by the L1 loss score [35].

(ii) Multi-Tasking SSL: To maximize the SwinT encoder's feature-capturing capabilities, we subsequently implemented multi-tasking SSL by assigning multiple challenging pretext tasks to the encoder. In addition to image inpainting, image rotation prediction and contrastive learning were introduced as suggested in the SSL SwinT paper [34]. Each input image was first duplicated to create a positive image pair for contrastive learning. Each duplicate was augmented differently by undergoing random masking and rotation by 0^o, 90^o, 180^o or 270^o, thereby enabling both image inpainting and image rotation prediction as a 4-class classification task.

Following the aforementioned encoding steps, three SSL heads were attached to the encoder in parallel. Image reconstruction was done via the same CNN decoder structure used in single-tasking SSL. A classification head was attached to predict the softmax probabilities for the 4-class rotation classification, guided by cross-entropy loss. A linear layer was attached to map each augmented image to a latent representation, using cosine similarity to measure the distance between positive and negative image pairs. The SSL performance of this model was measured using a combined loss function outlined in Table 1.

Table 1: Loss functions for the pretext tasks used in SSL pretraining [34].

Pretext Task	**Loss Function**	**Equation**
Image Inpainting	L1 loss	$L_{inpaint} = \|Y - \hat{Y}\|_1$
Image Rotation Prediction	Cross-entropy loss	$L_{rot} = -\sum_{r=1}^{R} y_r log(\hat{y}_r)$, where R is the number of class
Contrastive Learning	Contrastive loss	$L_{contrast} = -log\left(\frac{exp(2(v_i \cdot v_j))}{\sum_k^{2N} 1_{k \neq i} exp(2(v_i \cdot v_k))}\right)$, where v_i and v_j are latent representations of positive pair
Multi-tasking SSL	Combined loss	$L_{SSL} = \lambda_1 L_{inpaint} + \lambda_2 L_{rot} + \lambda_3 L_{contrast}$, $\lambda_1 = \lambda_2 = \lambda_3 = 1$

2.2.2 Stage 2: Supervised Downstream Segmentation

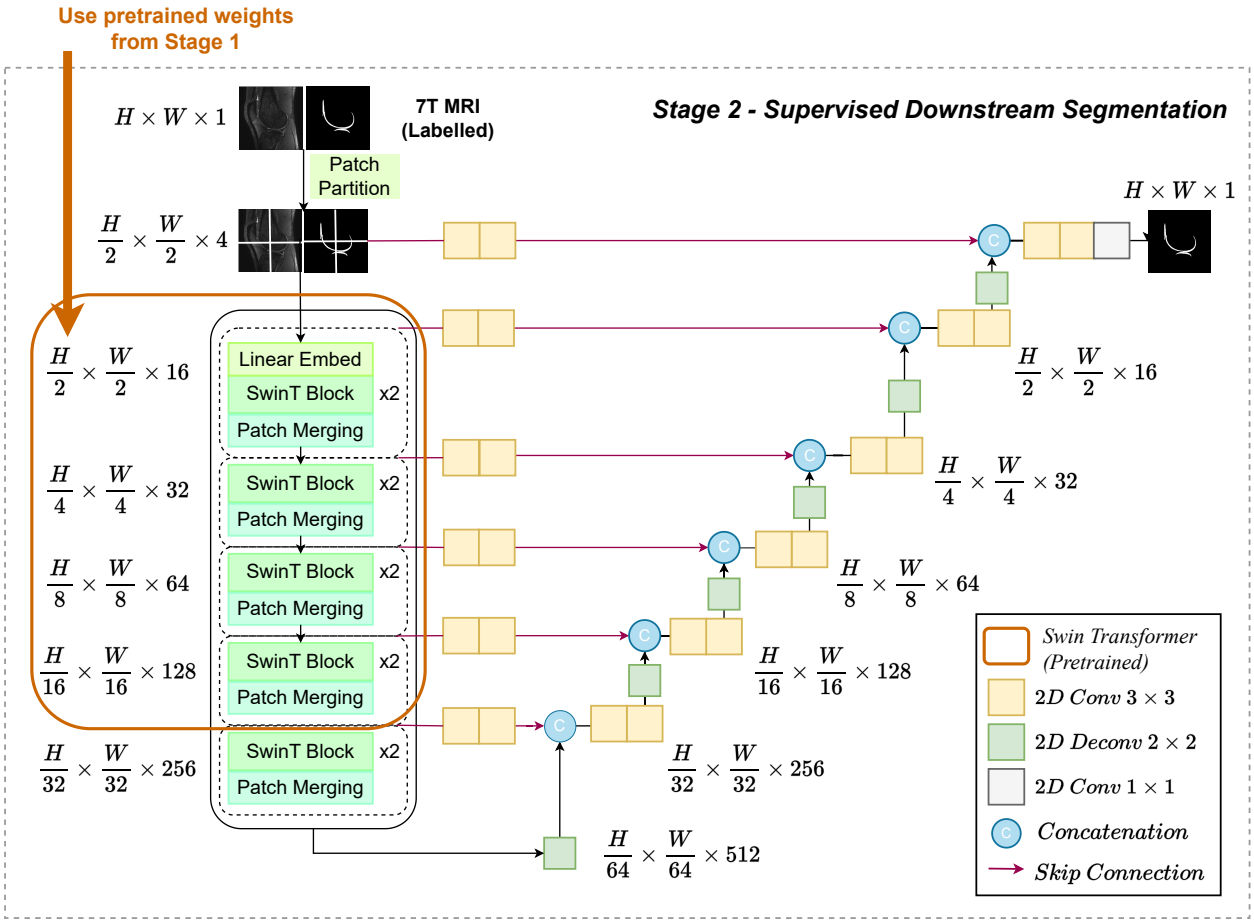


Figure 5: Illustration of the process and network architecture used in Stage 2 - Supervised Downstream Segmentation.

Figure 5 illustrates the supervised downstream segmentation of 7T MRI, also known as the fine-tuning stage. It utilized the 2D version of the Swin-UNETR architecture [34]. The U-shape segmentation network [27] consists of the SwinT encoder connecting to a CNN-based decoder through skip connections. Pretrained SSL weights from Stage 1 were loaded onto the encoder for knowledge transfer. Additional SwinT blocks were incorporated to accommodate the larger image dimensions of 7T MRI. During supervised training, image-label pairs were fed into the model, following the same encoding steps described in Stage 1. The CNN-based decoder, dimensionally symmetrical to the encoder, performed upsampling through transposed convolutional layers to progressively restore the image's spatial resolution. Skip connections were introduced to pass previously retained features from the encoder at each layer. A final 1×1 convolutional layer with sigmoid activation was employed to compute the segmentation probabilities ranging from 0 to 1. Using a threshold of 0.5, pixel values greater than 0.5 were assigned with a label of 1 to represent cartilage, and vice versa for background pixels.

2.3 Experiments

2.3.1 Dataset and Preprocessing

(i) Osteoarthritis Initiative (OAI) Database (3T-MRI): We utilized the OAI database [36], consisting of 3T MRI scans from an 11-year longitudinal cohort study. The database includes data from 4796 individuals, aged 45 to 79 years old, who are equally represented across gender and have either been diagnosed with or are at risk of femoral-tibial KOA. All knee MRIs are in DICOM format with 3T resolution. The majority of the MRIs are unannotated, except for 148 MRIs that were manually segmented by professionals from OAI for the Osteoarthritis Imaging Knee MRI Segmentation Challenge (OAI Challenge dataset) [37]. We used the unannotated OAI database for SSL pretraining and the OAI Challenge dataset as an initial trial and evaluation of the downstream segmentation task.

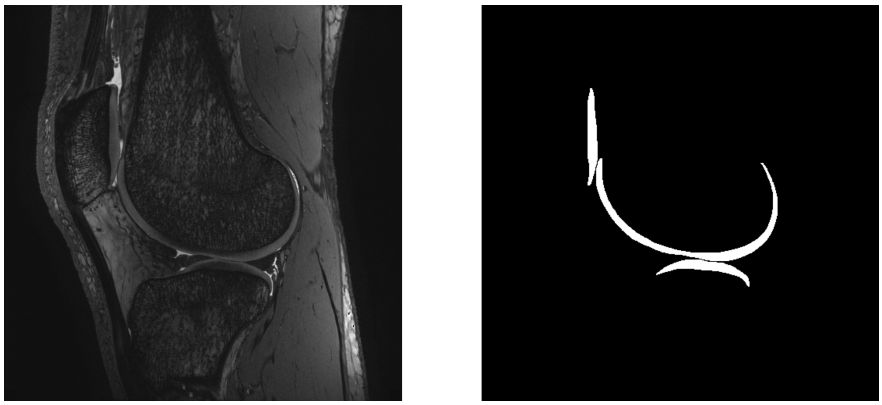
Each MRI volume has a shape of $384 \times 384 \times 160$. To convert into a 2D problem, each 3D volume is sagittally sliced into 160 2D images of a size of 384×384. Each image is downsized to 256×256 with average pooling to further reduce computational costs.

For the annotated OAI Challenge dataset, each image includes a 6-class segmentation label, including the femoral cartilage, medial and lateral tibial cartilages, patellar cartilage, and lateral and medial menisci. For simplicity, we combined all cartilage compartments to perform single-class segmentation only and discarded menisci labels.

(ii) 7T-MRI Dataset: The target 7T MRI dataset for downstream segmentation used in this study was acquired by the Bangerter Group at Imperial College London, in collaboration with the Department of Orthopaedics, University of Oxford. It consists of 14 3D knee MRI scans in 7T resolution from 8 subjects, comprising 8 men and 2 women. Each MRI volume has a dimension of $512 \times 512 \times 80$. Images are stored in DICOM format and were also sliced along the sagittal plane to obtain 80 2D images, each of size 512×512, with single-class cartilage labels.

All image slices are normalized to [0, 1] and converted into HDF5 format for more efficient storage and data access for neural network training. Figures 6a and 6b show the preprocessed annotated OAI 3T-MRI and 7T-MRI sagittal slices respectively.

(a) 3T MRI



(b) 7T MRI

Figure 6: Example of the knee MRI slice in the sagittal plane and its corresponding cartilage segmentation mask from (a) the 3T OAI Challenge dataset; (b) the 7T MRI dataset

2.3.2 Dataset Extraction

It is observed that among all slices from each MRI volume, there is a significant number of slices with very little or no cartilage that does not provide meaningful information. To ensure dataset representativeness and balance, we selectively extracted slices containing the most cartilage pixels as follows:

- **Unlabelled SSL Data**: Without cartilage annotation, a manual investigation of the MRI was performed to locate centres of medial and lateral condyles covered by the largest amount of cartilage sagittally. Due to the consistent knee distribution in every MRI across the dataset, we approximated slices 40-49 and 100-109 in every MRI to be the medial and lateral condyles centres, hence extracting 20 slices from each volume.

- **Labelled Segmentation Data**: With labelling, we located the slice number containing the most cartilage pixels in its corresponding label, selecting it and the consecutive 4 slices before and 5 slices after it, in total obtaining 10 slices from each MRI, either from the medial or lateral condyle.

After image extraction for both SSL pretraining and downstream segmentation datasets, we randomly shuffled and split each of them to obtain the training, validation and testing datasets for each stage, with details shown in Table 2.

Table 2: Train-valid split of the unlabelled SSL data and train-valid-test split of the 3T and 7T labeled downstream segmentation data. For unlabelled SSL data, the train-valid ratio is 0.8:0.2. For labelled segmentation data, the train-test ratio is 0.9:0.1, with 0.2 of the training dataset randomly assigned to be the validation dataset.

	Unlabelled SSL Data (3T OAI Database)	**Labelled Segmentation Data**	
		3T OAI Challenge	**7T MRI**
Train	4560	1065	100
Validation	1160	267	26
Test	N/A	148	14
Total	**5720**	**1480**	**140**

2.3.3 Implementation Details

For all trainings, the Adam optimizer was used for hyperparameter optimization, with an initial learning rate of 1e-3 and weight decay of 1e-5. For SSL pretraining, we used a batch size of 32 for the 3T OAI database. Specific loss scores used are outlined in Table 1.

During downstream segmentation, we first evaluated our method on the 3T OAI Challenge dataset using the same batch size in SSL pretraining. It was followed by the target 7T MRI segmentation, in which we adjusted the batch size to 8 for higher memory efficiency in higher-resolution images and better generalization in a smaller dataset. DiceCELoss [38], a weighted sum of Dice loss and Cross Entropy Loss, was used as the loss function. A 5-fold cross-validation strategy was implemented to mitigate potential bias in the training and validation dataset, leading to more representative and generalized results. Details about model parameters and training hyperparameters can be found in Appendix A.1.

Computations and neural network implementations in this study were performed using Python, utilizing the PyTorch framework (version 1.8.0). Codes for SwinT, Swin-UNETR and evaluation metrics were adapted from the open-source codes available in MONAI [39]. All models were trained on a GPU implementation with NVIDIA RTX 6000.

2.3.4 Evaluation

To evaluate the performance of downstream segmentation, two common metrics were used, Dice similarity coefficient (DSC) and Intersection-over-Union (IoU). DSC (Equation 2.1) is an overlapping spatial metric ranging from 0 to 1, indicating from no spatial overlap to complete overlap between segmentation prediction and ground truth, disregarding background pixels.

$$DSC = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{2.1}$$

where Y and $\hat{Y}$ represent the ground truth and prediction respectively. While using DSC might cause misleading results as the background dominates a significant portion of the image and ROI is small, IoU (Equation 2.2) considers both the overlapping and non-overlapping regions, penalizing under- and over-segmentation more than DSC [40].

$$IoU = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \tag{2.2}$$

Using both metrics ensures a fair assessment of the model's performance. Visual performance comparisons between predicted masks and ground truths are also crucial to ensure the alignment between quantitative and qualitative results.

Notably, the primary focus of this research was not to evaluate the results of SSL pretraining. We assessed its performance through qualitative analysis of reconstructed images and observation of loss function curves. Quantitative evaluation metrics were not utilized in this context.

2.3.5 Ablation Study

To investigate the significance of various components on the model's performance, we conducted experiments in the form of an ablation study. It is a systematic exploration where an element of the input is replaced or eliminated each time, aiming to observe the impact of each element on the downstream segmentation performance. Details of the ablation study plan are described in Table 3. Initially, we examined the effectiveness of SSL in Experiments 1-3. Subsequently, in Experiments 4-5, we evaluated the data efficiency of the model by conducting 10 sub-trainings with varying amounts of labelled fine-tuning data (ranging from 10% to 100%), with and without SSL. This aimed to determine whether the model could maintain accurate performance while minimizing annotation effort with the help of SSL.

Table 3: Ablation study plan. Experiments 1-3 compared the downstream segmentation performance upon no pretraining (purely supervised), single-tasked SSL pretraining and multi-tasked SSL pretraining. Experiments 4-5 involved 10 sub-trainings each (with no pretraining and multi-tasking SSL pretraining), from using 10% to 100% of the fine-tuning labeled dataset for data efficiency investigation.

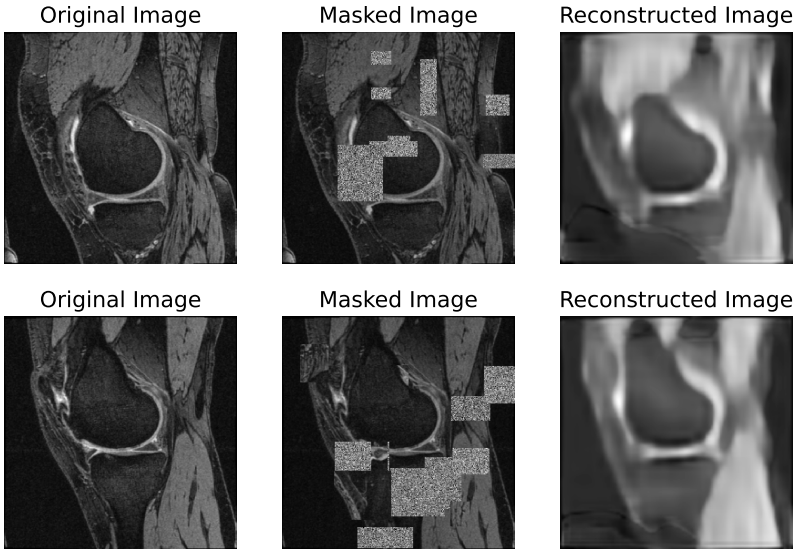
Experiment	SSL Pretraining	SSL Method	Percentage of labelled fine-tuning data used
1	No	N/A	100%
2	Yes	Single-tasking	100%
3	Yes	Multi-tasking	100%
4 (10 trainings)	No	N/A	from 10% to 100%
5 (10 trainings)	Yes	Multi-tasking	from 10% to 100%

3 Results

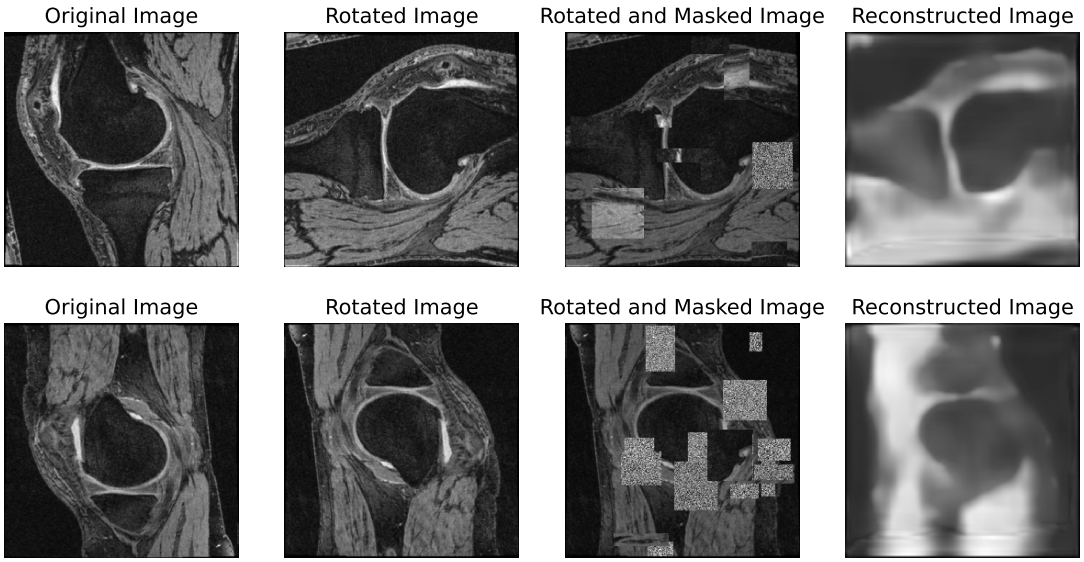
3.1 Stage 1: SSL Pretraining

Reconstructed images of the augmented image input from the 3T OAI database were obtained from both single-tasking SSL and multi-tasking SSL. Figure 7 shows two examples

from each SSL approach. Despite the slight blurriness, the shape of the knee and femur was clearly shown in the reconstructed image, with contrast and contours within the masked area correctly reconstructed. The blurriness might be attributed to the lower resolution of 3T MRI, resulting in difficulty in clearly distinguishing fine tissue structural details. Yet, with this level of reconstruction, this model outperformed a previous study [41] on the same dataset with SSL image inpainting task. This demonstrates its ability to learn informative feature representations from knee MRI through image inpainting.



(a) Single-tasking SSL



(b) Multi-tasking SSL

Figure 7: Two 3T MRI examples of (a) Single-tasking SSL: the original image was randomly masked and fed into the model for image inpainting, which was subsequently reconstructed; (b) Multi-tasking SSL: each original image was randomly rotated (by 90^o and 180^o respectively in the examples) and masked, which was subsequently reconstructed.

In multi-tasking SSL, the model has to complete three pretext tasks in parallel, therefore

exhibiting a slight decrease in the inpainting performance with less clarity in reconstructed images. However, this did not necessarily indicate a reduction in the model's ability to capture image features, as it still benefited from other pretext tasks. To visualize model performances on all tasks, Figure 8 displays the validation loss function curves for all three tasks. All loss curves converged to a minimum, indicating the model's progressive performance improvements across epochs by minimizing the discrepancy between predicted and actual values, signifying effective learning of all tasks. Particularly, the cross-entropy loss converged to approximately 0.01, indicating the high accuracy of the model in predicting the rotation degree of the image. However, among all loss functions, the contrastive loss remained significantly higher at approximately 2.7, suggesting challenges in optimizing the contrastive learning task. This concludes that the model successfully captured image feature representations from predictive (rotation classification) and generative (image inpainting) tasks, while not necessarily benefiting from contrastive learning. Detailed SSL pretraining results are shown in Appendix A.2.

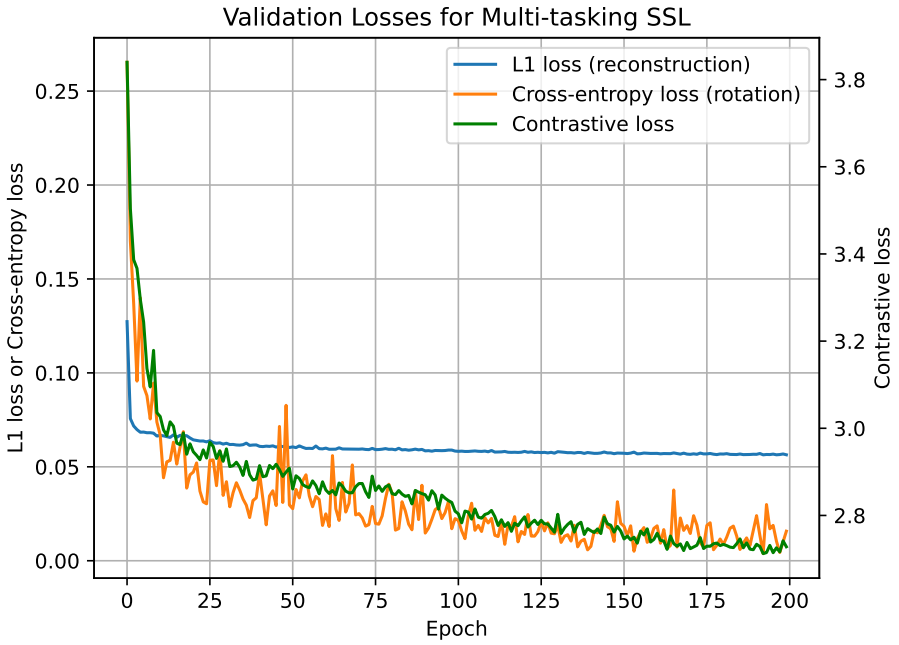


Figure 8: Validation loss curves (reconstruction L1 loss, rotation cross-entropy loss and contrastive loss) of the three tasks in multi-tasking SSL

3.2 Stage 2: Supervised Downstream Segmentation

3.2.1 Segmentation Results and Effectiveness of SSL

Supervised downstream segmentation was performed on the two labelled datasets we had with 5-fold cross-validation, first with the 3T OAI Challenge dataset, followed by the 7T MRI dataset. To assess the effect of SSL pretraining, we performed Experiments 1-3 in the ablation study, and model performances were compared using the same 3T and 7T test datasets across all experiments. Averaged DSCs and IoUs of the test predictions are summarized in Table 4.

Table 4: Segmentation performance (DSCs and IoUs) on the 3T OAI Challenge test dataset and 7T MRI test dataset in Experiments 1-3. Metrics were averaged across all test predictions with standard deviation (SD) calculated.

Experiment	3T OAI Challenge Dataset		7T MRI Dataset	
	DSC $(mean \pm SD)$	**IoU** $(mean \pm SD)$	**DSC** $(mean \pm SD)$	**IoU** $(mean \pm SD)$
1) No SSL	0.888±0.0035	0.800±0.0054	0.928±0.0040	0.867±0.0070
2) Single-tasking SSL	0.902±0.0033	0.823±0.0055	0.932±0.0010	0.872±0.0017
3) Multi-tasking SSL	$\mathbf{0.903 \pm 0.0007}$	$\mathbf{0.825 \pm 0.0012}$	$\mathbf{0.938 \pm 0.0011}$	$\mathbf{0.884 \pm 0.0019}$

Firstly, Table 4 shows that DSC generally exhibited higher values compared to IoU, as IoU imposes stricter penalties in segmentation. Secondly, for predictions of both datasets, accurate performance with high DSC and IoU was already achieved by the purely supervised model (no SSL), with slight improvement given by SSL pretraining. For the 3T OAI Challenge dataset, there is a further 1.58% DSC increase with single-tasking SSL pretraining and a 1.69% increase with multi-tasking SSL pretraining, reaching a DSC of 0.903. Comparatively, improvements in the 7T MRI segmentation were modest, with only a 0.43% and 1.08% increase with single-tasking and multi-tasking SSL pretraining respectively, reaching the highest DSC of 0.938.

Notably, for all Experiments 1-3, higher evaluation scores were obtained on the 7T MRI segmentation predictions compared to that of the 3T MRI. This reflects the model's ability in predicting well-segmented cartilage labels close to the ground truths especially in higher-resolution images. Lastly, standard deviations (SD) of both DSC and IoU across all experiments were insignificant (<0.01), indicating a remarkably consistent performance in predictions across the whole test dataset.

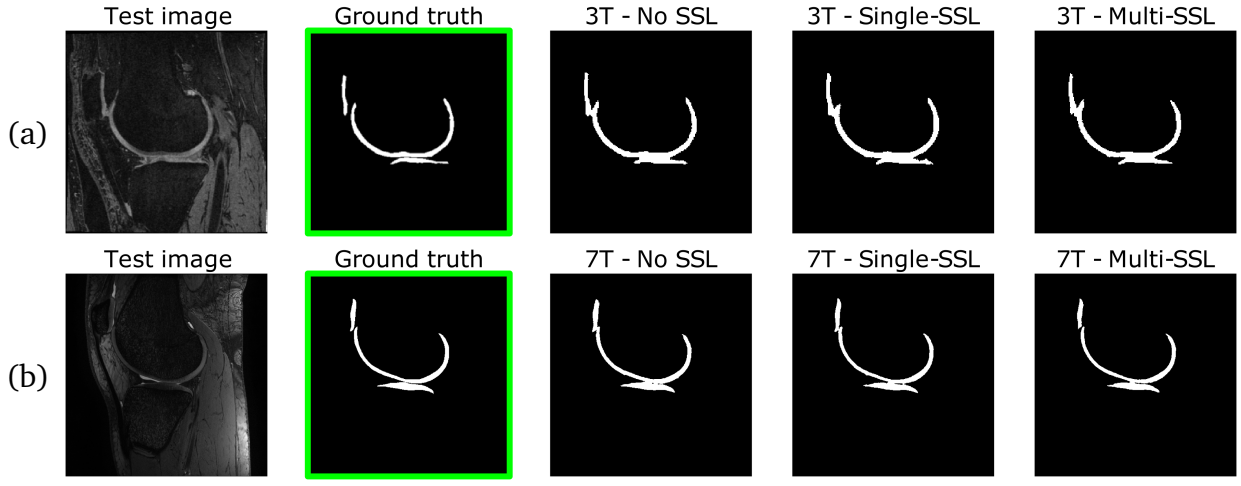


Figure 9: (a) 3T MRI test image (b) 7T MRI test image: examples that show no significant difference between the ground truth label, segmentation predictions from the purely supervised model (Experiment 1), models pretrained with single-tasking SSL (Experiment 2) and multi-tasking SSL (Experiment 3).

In terms of visual performance, for both 3T and 7T MRIs, due to the small difference in DSC and IoU across all three models, the majority of predictions from the three models did

not exhibit significant difference, with two examples shown in Figure 9. Aligning with the high DSC and IoU obtained, all predictions are visually similar to the ground truth, with the 7T predictions in general having smoother cartilage boundaries and higher accuracies.

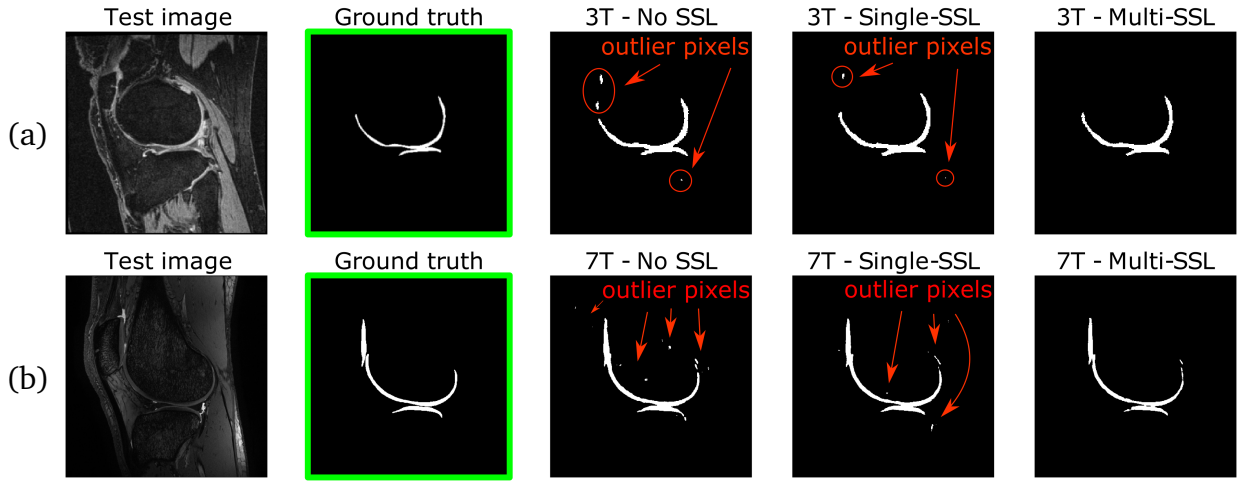


Figure 10: (a) 3T MRI test image (b) 7T MRI test image: examples with significant outlier pixels presented in the predictions from purely supervised model, which have been successively eliminated when using single-tasking SSL- and multi-tasking SSL-pretrained models.

Nevertheless, SSL pretraining did offer noticeable improvements in some predictions that were not well performed by the purely supervised model. Figure 10 highlights two cases, each from the 3T and 7T MRI where significant defects and outliers pixels were present in the purely supervised predictions, but were notably improved with single-tasking SSL, and mostly eliminated with multi-tasking SSL. This demonstrates the effectiveness of SSL pretraining in removing defects and providing almost outlier-free segmentations.

3.2.2 Data Efficiency

To evaluate the data efficiency of the model, Experiments 4-5 of the ablation study were conducted. Due to the limited size of the 7T MRI dataset (100 training images), 10% of the data referred to merely 10 images. Such a small dataset was insufficient to yield reliable and representative results for this investigation. Consequently, experiments were only performed on the larger 3T OAI Challenge dataset.

Figure 11 presents the best DSCs and IoUs during model validation obtained from both the purely supervised model and the multi-tasking SSL-pretrained model, varying the percentage of labelled fine-tuning data used. Both models exhibited suboptimal performance when being trained with less than 50% of the labelled data. Yet, in general, the SSL-pretrained model consistently outperformed the purely supervised model, even when utilizing only 10% of the data. The DSC and IoU of the SSL-pretrained model plateaued at approximately 0.91 and 0.83 respectively when 60% of the data was used. In contrast, the purely supervised model failed to converge to stable performance, with DSC and IoU fluctuating between 0.87-0.90 and 0.78-0.82 respectively. These findings highlight the consistent and accurate segmentation results achieved with the aid of SSL pretraining, offering a potential reduction in annotation effort by at least 40% on the 3T OAI Challenge dataset.

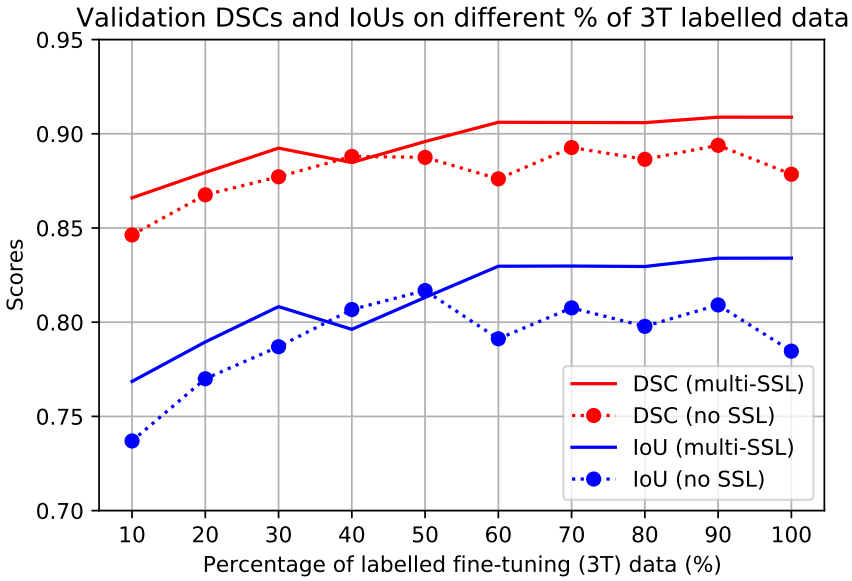


Figure 11: Data efficiency performance: validation DSCs and IoUs plots of both the purely supervised model (no SSL) and the multi-tasking SSL-pretrained model on different percentages of labelled 3T MRI dataset used for training.

4 Discussion

4.1 Segmentation Performance of the Proposed Method

We introduced the approach of combining the SwinT-based U-Net model and SSL pretraining on unlabelled 3T MRI for automated 2D knee cartilage segmentation of the limited labelled 7T MRI. Our methodology was first evaluated on the labelled 3T OAI Challenge dataset, serving as an initial indication of the model performance. With the purely supervised model trained on the extracted dataset with images containing large cartilage ROI, accurate 3T segmentation predictions were already obtained. The large ROI allows the model to extract features from broader image surroundings, enhancing its ability to capture precise boundaries and variations around the ROI. This mitigates small-scale variations and leads to more accurate segmentation predictions. Further leveraging pretrained weights from multi-tasking SSL, we achieved further refinements in segmentation in outlier removal, surpassing a DSC of 0.9.

Table 5 compares the 3T segmentation results of our proposed method with two top-performing 2D segmentation methods from the OAI Knee MRI Segmentation Challenge [37], using the same 3T MRI dataset. Our purely supervised model outperformed the two CNN-based U-Nets, and the addition of multi-tasking SSL pretraining further enhanced the performance. It is important to note that our method used a smaller dataset focusing on MRI slices with high cartilage content, whereas the published approaches used the complete set of MRI. While this comparison may lack complete fairness, it highlights the potential of our SwinT-based U-Net framework to achieve higher segmentation accuracy, even with limited data availability.

Table 5: DSC comparison of the 3T MRI segmentation between the proposed method in this paper and two methods published by the OAI Knee MRI Segmentation Challenge paper [37], using the same dataset. Both published methods performed multi-class cartilage segmentation, hence averaged DSCs across three cartilage compartments are reported below.

Network Achitecture	Training Method	DSC
2D CNN-based U-Net with multi-planar image sampling [37]	Supervised	0.880 (Averaged)
2D CNN-based U-Net with slices in sagittal plane [37]	Supervised	0.880 (Averaged)
2D SwinT-UNETR (the proposed method)	Supervised	0.888
	SSL pretrained + supervised fine-tuning	**0.903**

Following the 3T MRI segmentation, we performed training and prediction on our target 7T MRI dataset. Results of the 7T predictions followed a similar trend to the 3T predictions, with an overall higher DSC and IoU, as well as more accurate visual segmentation labels. This is due to the increased level of details and higher signal-to-noise ratio in ultra-high resolution MRI, providing more fine-grained details and clearer boundaries between tissues, facilitating a more precise extraction of boundary pixels around the ROI. Leveraging the pretrained knowledge from multi-tasking SSL, the model significantly reduced the outlier pixels present in predictions generated by the supervised model, achieving segmentation closely approaching the gold standard.

4.2 Limitation of the Proposed Method - Overfitting

Despite the high segmentation accuracy achieved, there are limitations to the proposed method. In the 7T segmentation results, the difference in DSC between the purely supervised model (0.928) and the two SSL-pretrained models (0.932 and 0.938) is insignificant. It could potentially be attributed to model overfitting to the small 7T dataset with low variability. Our focus on extracting MRI slices containing a high proportion of cartilage resulted in a training dataset consisting of only 100 similar images, leading to a lack of data variability that may result in model overfitting. The model quickly reached a performance plateau, impeding further improvement and limiting the potential benefits that could be derived from SSL pretraining. Although the model achieved highly accurate segmentation predictions on the test dataset, which had a strong resemblance to the training dataset, its performance might be suboptimal to unseen data. Regarding this, a preliminary investigation was conducted, as described in Appendix A.3, indicating a potential risk of model overfitting. This highlights the need of addressing challenges associated with training on small and low-variability datasets in the future, to optimize model generalizability and effectiveness of knowledge transfer from SSL pretraining to downstream segmentation. Future investigation can also focus on training the model with a diverse and balanced dataset that extends to all knee MRI slices, to improve the model's robustness in supporting the practical application of whole-knee cartilage segmentation.

4.3 Effectiveness of SSL Pretraining

From the SSL pretraining results presented in Section 3.1, the well-reconstructed images with clear contrast and contours from both single-tasking and multi-tasking SSL showed the model's ability to capture and learn useful image features for image regeneration from the inpainting task. In particular for multi-tasking SSL, the model had to learn surrounding pixel patterns for reconstruction and structural information for image rotation at the same time. It successfully captured both image structure and relative positioning of tissues and bones, which could be beneficial to downstream segmentation. Although the multi-tasking SSL model showed satisfying learning performances in inpainting and rotation classification tasks, it encountered difficulties in contrastive learning to differentiate between similar and dissimilar image pairs. In fact, this limitation can be attributed to the high consistency and similarity across all pretraining images, lacking diverse objects and structures necessary for effective contrastive learning. Considering this, contrastive learning may not provide useful insights for the downstream task and can be omitted.

Regarding the effectiveness of SSL pretraining for improving downstream segmentation, given the high prediction accuracy obtained from the purely supervised model, SSL pretraining provided a slight improvement for both 3T and 7T segmentations. While the purely supervised model was already capable of producing labels accurately presenting cartilage contours, SSL pretraining was effective in label refinement. Single-tasking SSL model eliminated some outliers, and the multi-tasking SSL model produced nearly defect-free segmentations. In medical imaging, achieving exceptional accuracy is paramount for practical implementation of DL applications. Therefore, despite the significantly longer training time and computational requirements needed, it is deemed worthwhile to opt for multi-tasking SSL over single-tasking SSL, leading to more accurate segmentation outcomes, aligning with the gold standard achieved by human annotation.

Overall, this SSL pretraining approach is considered effective in refining downstream segmentation predictions by leveraging large unlabelled MRI, particularly in removing outliers and artifacts. Several future improvements can be considered to maximize the effectiveness of SSL in downstream segmentation. Firstly, the unlabelled 3T MRI had a lower resolution of 256×256 in the experiment, whereas the target 7T MRI had a size of 512×512. As a result, in the downstream U-Net, additional un-pretrained encoder blocks were added to address the resolution gap between the two datasets. This might potentially constrain the effectiveness of knowledge transfer from low-resolution to high-resolution images. Regarding this, future implementation can consider upsizing the pretrained 3T MRI to match the dimension of the target 7T MRI using simple interpolation methods. Secondly, extending the pretraining dataset to include slices from the whole knee can potentially address the overfitting problem mentioned in Section 4.2, enabling the model to learn diverse feature representations and improve model generalization. Lastly, a thorough ablation study can be conducted to evaluate the effectiveness of individual pretext tasks and their combinations, to identify the most impactful approach for improving the downstream task. Exploration in parameters optimization of each pretext task such as the masking ratio for image inpainting can further enhance SSL performance.

4.4 Data Efficiency

Achieving data efficiency is another objective of this study, focusing on whether the model can maintain accurate and consistent performance through SSL pretraining with reduced annotated data. The multi-tasking SSL-pretrained model exhibited robust segmentation performance across various amounts of labelled data used, with higher accuracy compared to that of the purely supervised model. By leveraging SSL, the model effectively captured the underlying knee structure using pre-learned knowledge, reducing the reliance on the labelled data and leading to precise segmentation results, even with a reduced amount of labelled data. In contrast, the purely supervised model solely relied on the labelled data for training, making it more susceptible to dataset noise and variability, causing easier overfitting onto outliers or bias present in the dataset, hence the fluctuating accuracies across different labelled dataset sizes. Results concluded that the multi-tasking SSL-pretrained model demonstrated data efficiency by reducing annotation effort by 40% in the 3T MRI segmentation task. Based on the similarities observed between the trends in 3T and 7T segmentation results in Section 3.2.1, we can infer that comparable outcomes would be observed with a sufficient amount of 7T MRI data.

4.5 Future Work

In addition to the aforementioned future implementations, there are other potential areas of exploration for enhancing the SSL-pretrained MRI cartilage segmentation. Firstly, further investigation can be conducted to identify the most suitable combinations of auxiliary pretext tasks for MRI cartilage segmentation. This study focused on only three common pretext tasks used in the medical domain. Given the substantial differences between natural and medical images, domain-specific knowledge is necessary to extract meaningful features, particularly related to anatomical positioning of the ROI. For instance, Bai et al. [42] achieved improved segmentation accuracy in multi-class cardiac MRI by employing SSL with the anatomical position prediction task. Taleb et al. [43] investigated the performance of various pretext tasks in SSL for 3D MRI brain tumour segmentation, with the best DSC attained using relative patch location and jigsaw puzzle tasks. Combining the findings in this study, generative pretext tasks are considered more effective in extracting anatomical and structural information for segmentation. The proposed SSL pretraining framework is scalable and it supports multiple pretext tasks, allowing for easy implementation of other task combinations and augmentation methods in future research.

Secondly, the segmentation model can be extended to 3D MRI segmentation training, allowing the incorporation of spatial relationships across slices in real 3D MRI data. While this study focused on 2D training to reduce computational costs, considering the depth information of MRI is crucial for accurate cartilage volume quantification. Previous research has explored the use of 3D Transformer-based models, including SwinT [31] and other variations [44–49], in multi-organ MRI segmentations. These models have demonstrated increased accuracy and reduced computational and spatial complexities compared to pure 3D CNN-based methods. Therefore, 3D MRI segmentation with Transformer-based models can be a potential avenue for future work.

Additionally, this study focused on segmenting the cartilage as a single class. To provide a more comprehensive cartilage analysis, it would be beneficial to extend the model to support 4-class knee cartilage segmentation, differentiating between femoral, medial

tibial, lateral tibial and patellar cartilages. This would offer a more detailed representation for the knee joint, enhancing medical relevance for the use in thickness quantification.

Lastly, to support the assessment of KOA progression from cartilage segmentation results, cartilage thickness measurement should be considered as a model evaluation method in future investigation. Detailed calculation methodology is outlined in the OAI Challenge publication [37].

5 Conclusion

This study presented an automated cartilage segmentation framework for a limited amount of annotated 7T knee MRI. To address the problem of small dataset DL training, we leveraged the abundant unlabelled 3T knee MRI available for SSL pretraining on a SwinT model. The SwinT successfully learnt image feature representation from low-resolution MRI during SSL, especially in completing predictive and generative tasks. This allowed knowledge transfer to improve the downstream segmentation task, achieving accuracies that outperformed previous studies. The use of multi-tasking SSL-pretrained model further demonstrated the potential to achieve outlier-free segmentations, highlighting the possibility of reducing the need for manual annotation. Furthermore, our framework exhibited data efficiency, maintaining accurate performance even with a small amount of data, which is crucial for tackling challenges posed by small datasets. Future investigation should focus on extending the framework to enable whole-knee segmentation in 3D, while addressing the potential overfitting issue through enhancing image diversity, as well as adjusting corresponding SSL and fine-tuning strategies. Exploring other pretext tasks and multi-class segmentation can further optimize the model performance, ultimately facilitating the clinical application in cartilage thickness quantification and early-stage KOA detection.

References

[1] Hospital for Special Surgery (HSS) 2021. Knee osteoarthritis. https://www.hss.edu/condition-list_osteoarthritis.asp. [accessed: 2022-01-08].

[2] David J Hunter, Lyn March, and Mabel Chew. Osteoarthritis in 2020 and beyond: a lancet commission. *The Lancet*, 396(10264):1711–1712, 2020.

[3] Mark D Kohn, Adam A Sassoon, and Navin D Fernando. Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®*, 474:1886–1893, 2016.

[4] Chris Buckland-Wright. Which radiographic techniques should we use for research and clinical practice? *Best Practice & Research Clinical Rheumatology*, 20(1):39–55, 2006.

[5] Behzad Heidari. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part i. *Caspian journal of internal medicine*, 2(2):205, 2011.

[6] Russell C Fritz and Lynne S Steinbach. Magnetic resonance imaging of the musculoskeletal system: Part 3. the elbow. *Clinical Orthopaedics and Related Research®*, 324:321–339, 1996.

[7] F Eckstein, JE Collins, MC Nevitt, JA Lynch, V Kraus, JN Katz, E Losina, W Wirth, A Guermazi, FW Roemer, et al. Cartilage thickness change as an imaging biomarker of knee osteoarthritis progression–data from the fnih oa biomarkers consortium. *Arthritis & rheumatology (Hoboken, NJ)*, 67(12):3184, 2015.

[8] Dong Zhang, Yi Lin, Hao Chen, Zhuotao Tian, Xin Yang, Jinhui Tang, and Kwang Ting Cheng. Deep learning for medical image segmentation: tricks, challenges and future directions. *arXiv preprint arXiv:2209.10307*, 2022.

[9] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[10] M Akhil, R Aishwarya, Vandhana Lal, and Shanthi Mahesh. Comparison and evaluation of segmentation techniques for brain mri using gold standard. *Indian J Sci Technol*, 9(46):1–5, 2016.

[11] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J Humaidi, Omran Al-Shamma, Mohammed A Fadhel, Jinglan Zhang, J Santamaría, and Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7):1590, 2021.

[12] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Gokul Karthik Kumar, Sahal Shaji Mullappilly, and Abhishek Singh Gehlot. An empirical study of self-supervised learning approaches for object detection with transformers. *arXiv preprint arXiv:2205.05543*, 2022.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*, 2021.

[18] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[19] Gayle Lester. The osteoarthritis initiative: a nih public–private partnership. *HSS journal*, 8(1):62–63, 2012.

[20] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[22] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016.

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[25] Saleh Albelwi. Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

[26] Tin Lam Cheng. Proc. project planning report. *Self-Supervised Learning for Cartilage Segmentation of Ultra- High Resolution Knee MRI*, 2022.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[28] Xiao-Xia Yin, Le Sun, Yuhan Fu, Ruiliang Lu, and Yanchun Zhang. U-net-based medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8354–8363, 2022.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[32] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022.

[33] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021.

[34] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022.

[35] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[36] National Institutes of Health. Osteoarthritis initiative (oai) study protocol. `https://nda.nih.gov/oai/study-details.html`[Accessed: 2023-05-15].

[37] Arjun D Desai, Francesco Caliva, Claudia Iriondo, Aliasghar Mortazi, Sachin Jambawalikar, Ulas Bagci, Mathias Perslev, Christian Igel, Erik B Dam, Sibaji Gaj, et al. The international workshop on osteoarthritis imaging knee mri segmentation challenge: a multi-institute evaluation and analysis framework on a standardized dataset. *Radiology: Artificial Intelligence*, 3(3):e200078, 2021.

[38] MONAI. Loss functions, loss functions - monai 1.2.0 documentation. `https://docs.monai.io/en/stable/losses.html`[accessed: 2023-06-05].

[39] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[40] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):1–8, 2022.

[41] Zirui Cao. Automatic knee cartilage segmentation based on a non-contrastive self-supervised learning method. *Unpublished manuscript*, 2022.

[42] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 541–549. Springer, 2019.

[43] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

[44] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023.

[45] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[46] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 78–88. Springer, 2021.

[47] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 109–119. Springer, 2021.

[48] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[49] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical image segmentation using squeeze-and-expansion transformers. *arXiv preprint arXiv:2105.09511*, 2021.

A Appendix

A.1 Model Parameters and Hyperparameters

Table A.1: Summarization of the dataset, network architectures, model parameters and hyperparameters used in SSL pretraining and downstream segmentation respectively.

	SSL Pretraining	**Downstream Segmentation**	
Dataset	3T OAI Database	3T OAI Challenge Dataset	7T MRI Dataset
Network Architecture	2D SwinT	2D SwinT-UNETR	
Maximum ROI dropping ratio in image inpainting	30%	N/A	
Embedding size in contrastive learning	256	N/A	
Encoder depth	4	4	5
Number of heads for MSA in each layer	[4, 8, 16, 32]	[4, 8, 16, 32]	[4, 8, 16, 32, 64]
Batch Size	32	32	8
Optimizer	Adam	Adam	
Learning Rate	1e-3	1e-3	
Weight decay	1e-5	1e-5	
Number of Epochs	200	100	150

A.2 SSL Pretraining Results

Figure A.1 shows the training and validation loss curves during single-tasking SSL pretraining, which is the L1 loss for the image inpainting task. Figure A.2 shows the training loss curves during multi-tasking SSL pretraining, including the L1 loss for image inpainting, cross-entropy loss for image rotation classification, and contrastive loss for contrastive learning. The corresponding validation loss curves are shown in Figure 8.

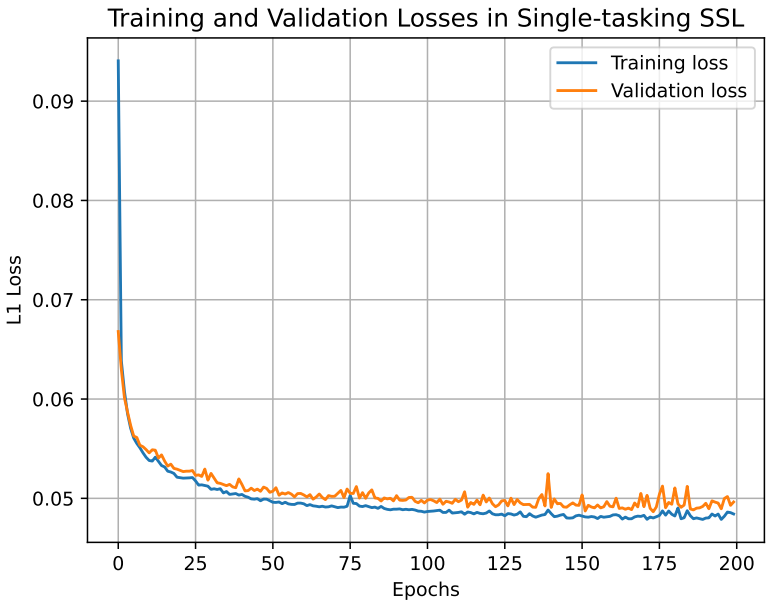


Figure A.1: Training and validation loss curves of single-tasking SSL.

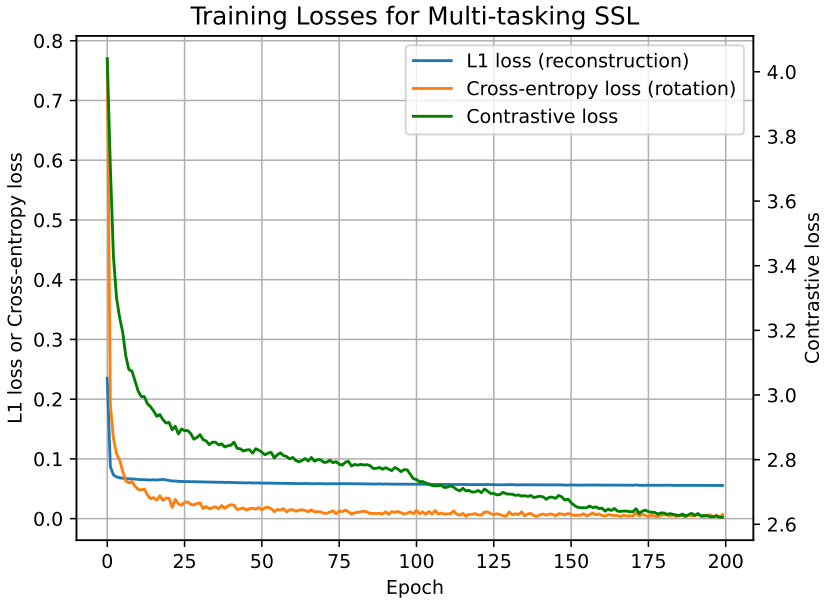


Figure A.2: Training loss curves of multi-tasking SSL.

A.3 Preliminary Investigation on Model Overfitting

To gain deeper insights into the overfitting problem potentially caused by the small 7T dataset with low variability between MRI slices with large cartilage, an additional purely supervised model training was conducted using all 7T MRI slices. The dataset primarily comprised sagittal cartilage-containing slices, with a small portion of slices near the medial and lateral ends of the knee that have minimal or no cartilage. There were in total 806 training, 202 validation and 112 testing images. The experiment yielded an average DSC of 0.909 and IoU of 0.837 across all test predictions, which had a reduction of performance by 2.05% compared to the purely supervised segmentation predictions on the extracted 7T dataset (0.928). Figure A.3 shows the visual segmentation results of this model. It was able to predict accurate segmentation for cartilage-containing slices. However, it encountered difficulties in generating segmentation masks for images containing little or no cartilage. This indicates that the model is easily overfitting to the dominant large-cartilage-containing slices in the 7T dataset. This highlighted the future investigation into maintaining model robustness in segmentation with non-uniformly distributed cartilage across images in the dataset.

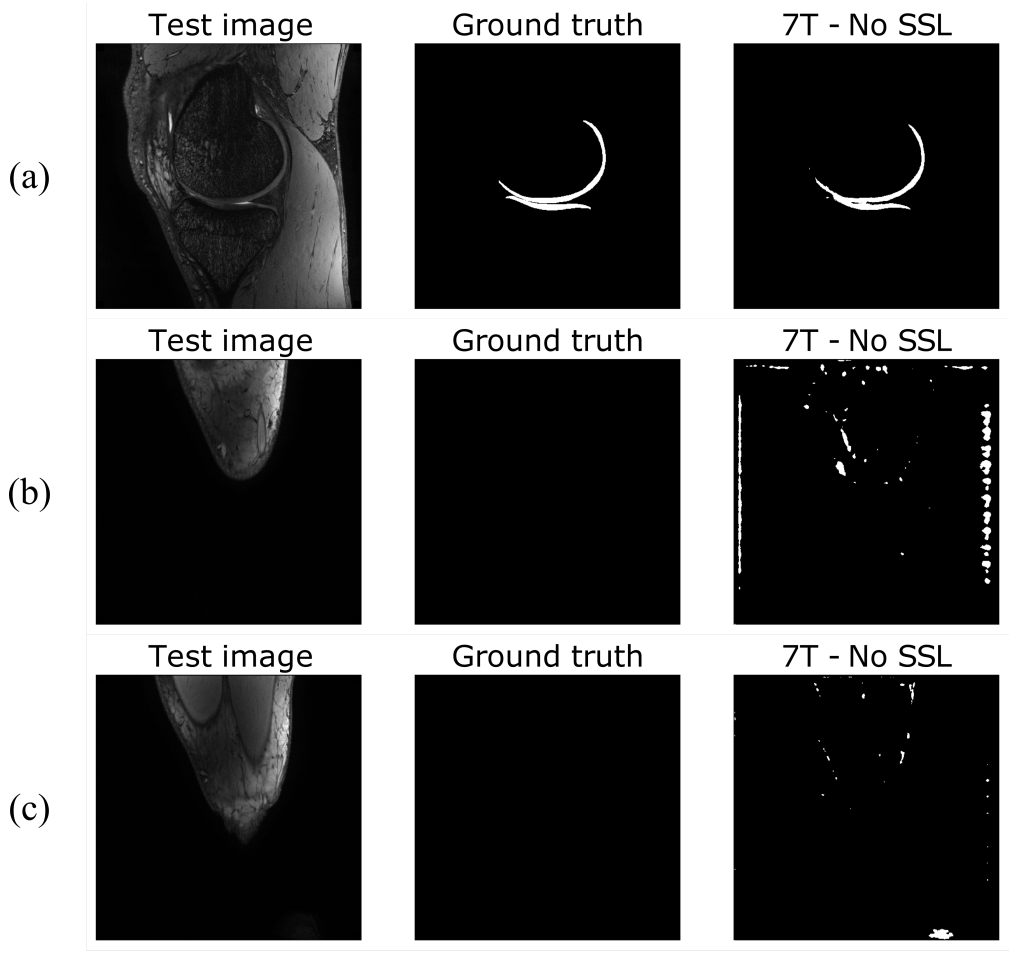


Figure A.3: Examples of the test image, its corresponding ground truth and segmentation prediction from the purely supervised model on all 7T MRI slices: (a) Well-segmented predictions on images with large cartilage regions; (b), (c) Inaccurate predictions from the model on images with no cartilage