Automated 3D Rat Vertebral Body Segmentation in μCT Images using Swin-UNETR

Seyedmohammadsaleh Mirzatabatabaei[*]
University of Toronto
s.mirzatabatabaei@mail.utoronto.ca

Sherry Yuan[†]
University of Toronto
shanli.yuan@mail.utoronto.ca

Alexander Ryabchenko[†]
University of Toronto
alexander.ryabchenko@mail.utoronto.ca

Abstract

A novel approach to improve rat vertebral segmentation employs Swin-UNETR in combination with Asymmetric Unified Focal loss with ratio constraint (AUFRC) and dilated attention mechanisms. Building on the success of Swin-UNETR in various medical imaging tasks, the proposed model improves performance of semantic segmentation of travecular bone within vertebrae in rat spines, enabling better comparison of biomarkers in cancer research and broader applications in medical imaging. The study contributes to the development of efficient image segmentation techniques, impacting diagnosis and treatment for patients with cancer and musculoskeletal conditions. The code is open sourced[3].

1 Introduction

Vertebral labelling and segmentation are essential components of an automated spine processing pipeline. These tasks play a crucial role in the investigation of the impact of cancer on bone and muscle quality and health. Accurate and reliable processing of rat spine micro-CT (μCT) scans can enable the comparison of biomarkers across both preclinical and retrospective data.

In recent years, Swin-UNETR [2] has demonstrated promising results on various 3D segmentation tasks, including brain tumour and organ segmentation. Although it have not been applied on the task of vertebral segmentation. The proposed research aims to leverage this success and apply it to the challenge of vertebral segmentation in rat spines. Currently, the most recent and relevant rat spine segmentation approach is using demons deformable registration followed by level set curvature evolution within the Amira software platform incorporating algorithms from the ITK toolkit [12]. This approach required the use of proprietary software and is not a fully automated process. By proposing Swin-UNETR with dilated mechanism and custom loss, we aim to contribute a high performance, fully automating vertebral segmentation that solely rely on open source tools. High performance and automated segmentation allows for robust and repeatable measurement of the effects of metastatic disease and treatments on vertebral bone, crucial to evaluating spinal stability. Swin-UNETR is a transformer-based model that is designed to process images as sequences of patches, allowing it to capture global context information. We believed Swin-UNETR will be advantageous at this task because maintaining a global context is important for the task of vertebral segmentation given the vertebrae tend to span the full height of a CT scan.

[*] PhD student and research assistant at Orthopaedic Biomechanics Lab - saleh.mirtaba@sri.utoronto.ca

[†] Visiting researchers at Orthopaedic Biomechanics Lab of Sunnybrook Research Institute.

[3] https://bitbucket.org/OrthopaedicBiomechanicsLab/vertebral-segmentation-rat-l2/src/master/

This investigation modified the Swin-UNETR for the vertebral segmentation task by including a dilated attention mechanism and training using a novel loos function based on the asymmetric unified focal loss [6]. The dilated mechanism aims to achieve wider context at each skip connection level and increase the field vision while reducing depth of the network. The modified loss function was introduced to address the inherent class imbalance in the segmentation process. Ablation showed that the modified loss and architecture improved the performance the Swin-UNETR model.

2 Related work

The Swin Transformer has emerged as a promising Vision Transformer that utilizes shifted windows to process images, as described in [1]. Swin-UNETR, presented in [2], combines the Swin Transformer with UNet, a popular model for image segmentation tasks. Moreover, [3] describes effective Swin Transformer pretraining procedures that can significantly reduce the amount of labelled training data required for training Vision Transformers.

D-Former, introduced in [4], is another attention mechanism that computes self-attention in a dilated manner, enlarging receptive fields and effectively capturing long-range information. Additionally, Unified Focal loss, discussed in [6], generalizes Dice and cross-entropy-based losses to handle class-imbalanced medical image segmentation, making it a promising solution for medical imaging applications.

There are many works that apply high-performing vision transformers and related models (UNet, Swin-Unet, Swin-UNETR) to the task of medical segmentation. These include but are not limited to [2; 3; 4; 9; 10; 11].

3 Methods

3.1 Data preprocessing and Augmentation

In this study, we focused on the automated segmentation of the L2 vertebral region in high-resolution μCT scans of rat spinal vertebral bones (Original T13-L4 μCT scan). The μCT images were sourced from six different rat groups, named series 700, 800, 900, 1000, 1100, and 2000, encompassing different ages and sizes. The scans were collected by research assistants between 2005 and 2022 and were performed using a μCT-100 device (Scanco Medical AG, Bruettisellen, Switzerland) located at the Orthopedic Biomechanics Laboratory (OBL) at the Sunnybrook Research Institute (SRI). Excised T13-L4 vertebral motion segments of each rat were μCT scanned adjacent to hydroxyapatite (HA) calibration phantoms, with scan parameters set at 55kVp, 200µA, 11W, and 34.4µm voxel spacing.

To prepare the algorithm ground truths, we segmented the whole bone volume within each L2 vertebra using a two-step process. First, we used a semi-automated atlas-based registration approach[12]. This process involved creating an atlas through the manual segmentation of a single spinal motion segment. Subsequently, the atlas scan was registered to the target vertebra using automated affine registration and demons deformable registration. The segmentation was then refined using the level set method to capture the entire vertebra. This process was conducted in the Amira 3D software (version 2021.2, Thermo Fisher Scientific). Next, calculated L2 vertebrae segmentations within each series were evaluated by orthopedic researchers at OBL. In some cases, the calculated segmentations were corrected manually within the 3D Slicer (version 5.2.1, `www.slicer.org`) , resulting in an accurate representation of the rat vertebrae labels.

Despite the scans being collected for various research purposes over the years, they were not consistent or suitable for deep learning applications. Therefore, we implemented several comprehensive preprocessing steps on the data to make them usable for our project.

The preprocessing can be divided into four phases:

1. Data Extraction and Format Conversion: We extracted the useful data from the OBL cloud at SRI and converted the raw μCT scan formats (DICOM series) to Amira format and then to a common NIfTI format for more convenient image processing. Images were cropped to focus on the L2 vertebral region, ensuring consistent bounding boxes (Figure 3). We utilized Slicer and the SimpleITK module to preserve the physical space features of the images. All images were extracted from unorganized data clouds and were named and organized clearly.

2. Resampling and Standardization: Both scans and segmentations were resampled to a sampling frequency of (0.035, 0.035, 0.035) to maintain consistent image spacing and standardize voxel sizes across all images. The scalar data types were corrected to 'int16' for scans and binary for segmentations. Image intensity was made consistent within the dataset by clipping scan data in the range of (-1000, 10000). Additionally, the scan and segmentation of a rat were corrected if their physical spaces, such as image origin, did not match.
3. Image Resizing: We resized segmentations and scans (image dimensions) without losing quality based on a scale factor to achieve consistent L2 vertebrae images across the dataset. The scale factor was chosen based on the computation of rat L2 vertebrae volume and our prior knowledge of rat sizes within each series.
4. Dataset Validation: In the final stage, the 700 and 900 series segmentations were excluded from the dataset as they did not pass our second validation check. Our final dataset consisted of 197 rat images for the pretraining stage, including 144 labeled and segmented images for fine-tuning. A sample scan is presented in Figure 4.

In order to increase the diversity of samples, several image augmentation techniques were applied:

1. Rotation and translation to increase the dataset's spatial variety.
2. Images are first resampled from (0.035, 0.035, 0.035) to (0.08, 0.08, 0.08) image spacing, then applied a random shift before importing for training.
3. Random cropping around ROI such that input images have consistent size.
4. Random intensity scaling that is at most 10% lower/higher than original image.

3.2 Model Architecture

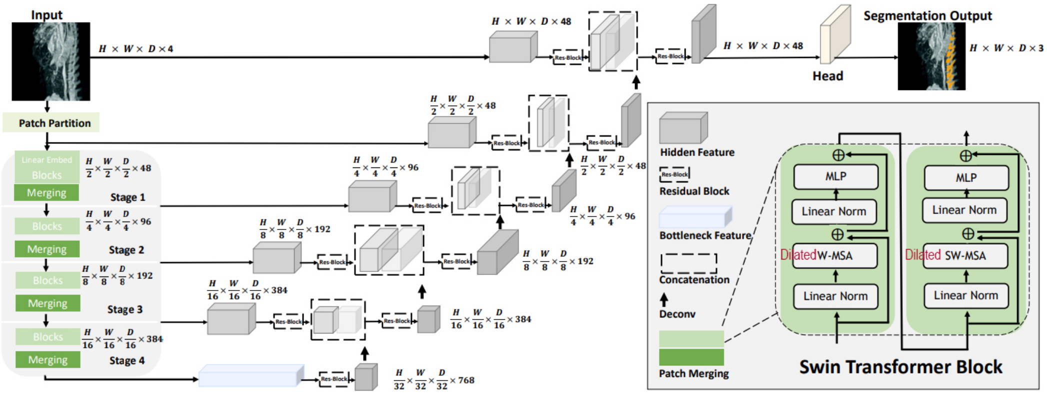
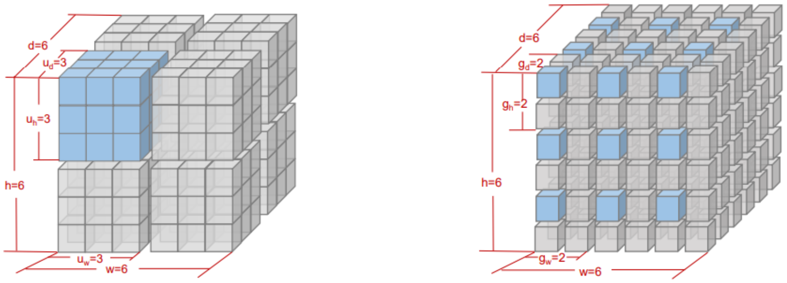

Input
$H \times W \times D \times 4$
Patch Partition
Linear Embed
Blocks
Merging
Stage 1
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
Blocks
Merging
Stage 2
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
Blocks
Merging
Stage 3
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
Blocks
Merging
Stage 4
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$
Res-Block
$H \times W \times D \times 48$
Head
Segmentation Output
$H \times W \times D \times 3$
Hidden Feature
Residual Block
Bottleneck Feature
Concatenation
Deconv
Patch Merging
MLP
Linear Norm
Dilated W-MSA
Dilated SW-MSA
Swin Transformer Block

Figure 1: Overview of Dilated Swin UNETR architecture (A modification to [2]). The input to our model is 3D multi-modal μCT images with 1 channel. The Dilated Swin UNETR creates non-overlapping patches of the input data and uses a patch partition layer to create windows with a desired size for computing the self-attention. The added dilation attention mechanism not compute the self-attention across every g-th voxel across multiple neighboring patches. The encoded feature representations in the Swin transformer are fed to a CNN-decoder via skip connection at multiple resolutions. Final segmentation output consists of 2 output channels corresponding to background and L2 vertebral. The model parameter size and FLOPs can be seen in Table 2 in Appendix A

The proposed model architecture modifies the vanilla Swin-UNETR [2] by adding a dilated attention mechanism to each patch merging stage of the Swin transformer. The model architecture is illustrated in Figure 1, and the dilation attention mechanism is illustrated in Figure 2.

The vanilla Swin transformer's approach of starting with small voxel patches and gradually merging them to form a final latent representation of an image is a beneficial strategy for the model to gain knowledge of both local and global contexts [1]. However, these benefits are less significant for large images, that require networks of great depth to maintain global attention.



(a) LS-MSA

(b) GS-MSA

Figure 2: (a) The local scope multi-head self-attention: voxels are adjacent. (b) The global scope multi-head self-attention (adapted from [4]): voxels are picked across the feature map with dilation.

The dilated mechanism (adapted from [4]) is implemented as follows. The self-attention is conducted in a global unit so that every g-th voxels (coloured blue in Figure 2b) are picked across the feature map with dilation. This happens before each voxel patch merge step of the Swin transformer. Then, the self-attention between these dilated voxels is computed to achieve wider context.

The global context is important for the task of vertebral segmentation because vertebrae tend to span the full height of a CT scan. The dilated attention mechanism allows self-attention to communicate with voxels outside of the local unit. This way, the merge process can be performed with fewer layers while maintaining the global context and computational efficiency. The final architectural diagram can be seen in Figure 1.

3.3 Pretraining

For pre-training tasks, the pipeline takes in 204 rat μCT scans, and learn the following tasks similar to [3]: (1) masked volume inpainting: the ROI dropping rate is set to 30 (as also used in [3]); the dropped regions are randomly generated and they sum up to reach overall number of voxels; (2) 3D contrastive coding: a feature size of 512 is used as the embedding size; (3) rotation prediction: the rotation degree is configured to 0, 90, 180, and 270. We train the model using the AdamW optimizer [13] with a warm-up cosine scheduler of 500 iterations. The model weights are initialized to those of the legacy model that was trained on 5050 public CT images from various human organs [3]. The pre-training experiments use a batch-size of 4 per GPU (with 96 x 96 x 96 patch), and initial learning rate of $6e-7$, momentum of 0.9 and decay of $1e-5$ for 6000 iterations. Our model is implemented in PyTorch and MONAI4, and it was executed on Compute Canada Mist Cluster in 24 hours using single NVIDIA V100-SMX2-32GB GPU. A five-fold cross validation strategy is used to train models for all experiments. We select the best model in each fold and ensemble their outputs for final segmentation predictions.

Note the population of rats in input images is not uniform, e.g. rats may be of different age. This variation in rat population is acceptable because the purpose of pretraining is to adapt the starting model (trained on CT image) to rat μCT images.

3.4 Finetunning

The pretrained models (with pretrain on rat μCT or only with pretrain on human organ CT) from above was then fine-tuned on a subset of 57 rat images labelled for L2-vertebrae segmentation. The constrastive learning, image inpainting and rotation prediction heads were removed and the extracted representations are then connected to a CNN-based decoder with skip connection. A segmentation head is attached at the end of the decoder for computing the final segmentation output (Figure 1). Ablation experiments were run to demonstrate the value of AUFRC (proposed in in Section 3.5) and pretraining.

Each fine-tuning task trained for 800 epochs using the AdamW, with learning rate $1e{-}4$. It was executed on Compute Canada Mist Cluster for 2 days using single NVIDIA V100-SMX2-32GB GPU.

3.5 Finetunning Loss function

The majority of models in medical image segmentation utilize cross-entropy loss, Dice loss, or some combination of both. Unified Focal loss [6] generalizes these approaches for handling class imbalance. Previously, weighted Tversky [7], focal Tversky [8], and mean square error of the bone volume fraction (BV/TV) proved to be successful in UNet-model for 3D segmentation [9].

In this paper, we propose a novel loss function - AUF with ratio constraint (AUFRC): weighted Asymmetric Unified Focal loss (AUF) [6] and mean square error of the bone volume fraction as ratio constraint (RC) (weights 2 and 0.5 respectively). Altogether, the loss function was as follows:

$$loss = W_{AUF} L_{AUF} + W_{RC} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{BV}{TV} - \frac{\widehat{BV}}{TV} \right)^2, \tag{1}$$

$$L_{AUF} = \lambda L_{AF} + (1-\lambda) L_{AFT}, \tag{2}$$

where W is the corresponding weight, L — corresponding loss function, BV and $\widehat{BV}$ are true and predicted bones volumes, and TV is total volume (bone and surrounding soft tissue). Here, AUF loss is taken directly from [6]. That is, AUF is a weighted Asymmetric Focal loss (AF) and Asymmetric Focal Tversky loss for parameters $(\lambda, \delta, \gamma)$. As per recommendations [6], parameters were set as $(\lambda, \delta) = (0.5, 0.6)$ and we take focal parameter $\gamma = 0.8$ for significant enhancement of the rare class.

For the problem of rat vertebral segmentation, AUF loss was chosen to address the inherent class imbalance in the problem, prevent harmful suppression of the rare class, and provide helpful balance between distribution-based Asymmetric Focal loss and region-based Asymmetric Focal Tversky loss. The constraint on bone-tissue volume ratio was introduced to further target the class imbalance.

4 Results

The results of the ablations are:

- Pretraining resulted in higher performance in most cases.
- AUFRC had the small edge in performance over the Dice loss.
- Dilated Swin-UNETR perform better than Swin-UNETR when there was no pretrain. However, Swin-UNETR has higher absolute performance in all metrics (considering both with and without pretrain).

The trained model consistently yielded high Dice Similarity Coefficient (DSC): 0.930 ± 0.002 and Intersection over Union (IoU): 0.870 ± 0.004, with strong agreement for BV/TV as trabecular bone microstructure parameter (0.976 ± 0.010). The best performing model is Swin-UNETR trained with proposed novel AUFRC loss. See Table 1 for training results. While differences were not statistically significant, the automated segmentation using AUFRC had the small edge in performance over the Dice Loss across all considered metrics.
The model trained with/without pretrain on μCT has little difference in DSC. In some instances, the pretrained model demonstrated even smaller accuracy. Moreover, we observed that the pretrained model has higher accuracy at the start of training, but as the number of epoches increases, the effect of pretrain becomes negligible.

We have also explored effects of increasing image spacing. With greater image spacing (0.08mm vs 0.035mm), the model's 96x96x96 cubes have greater view of the μCT scans. This yielded an increase in DSC of about $+0.08$.

5 Discussion

Utilizing this automated segmentation method yielded strong agreement with manual segmentation. The proposed model's performance is comparable to similar models for medical segmentation in a

	DSC		IoU		BV/TV R^2-value	
Pretrain	Yes	No	Yes	No	Yes	No
swinUnetr + Dice	0.9315	0.9312	0.8728	0.8722	0.9658	0.9703
swinUnetr + AUFRC	**0.9323**	0.9283	**0.8740**	0.8670	**0.9834**	**0.9855**
dilatedSwinUnetr + Dice	0.9313	0.9301	0.8722	0.8703	0.9774	0.9820
dilatedSwinUnetr + AUFRC	0.9304	**0.9322**	0.8708	**0.8737**	0.9571	0.9753

Table 1: Quantative results of model training. Here, swinUnetr stands for the original Swin-UNETR model without modifications, dilatedSwinUnetr stands for Swin-UNETR with dilated attention mechanism. Qualitative results can be seen in Figure 5.

3D setting (UNets [9; 10], Swin-Unet [11], Swin-UNETR [2]), with respect to considered metrics (Table 1).

Applying pretraining procedure, did not provide noticable increase in performance. This could be attributed to the insignificant amount of 200 rat μCT images used for pretraining, compared to 5000 human CTs used to train the original model.

It is important to note that the dilated Swin-UNETR was trained from scratch and did not reuse weights from the original model. So, dilated Swin-UNETR had to rely exclusively on 200 rat μCTs, which put it at a disadvantage when compared to non-dilated Swin-UNETR.

The performance of this method can be further improved with hyperparameter tuning: parameters of the AUFRC loss function - $(\lambda, \delta, \gamma, W_{RC})$ and learning rate. Due to time constraints, we were unable to fully explore the hyperparameter space. We believe that with more time and resources, we could have obtained better results. Future work could focus on performing a thorough tuning of hyperparameters to optimize the model's performance.

6 Conclusion

In conclusion, this study presents an automated method for segmenting rat vertebrae in μCT images using Swin-UNETR, a promising development for medical research related to vertebral diseases, disorders, and stereology. Our approach streamlines data analysis, offers more accurate and consistent results, and enables researchers to delve deeper into the underlying mechanisms of conditions like osteosarcopenia while exploring novel treatment options.

Additionally, we introduced the dilated Swin-UNETR, an innovative architecture for semantic segmentation that enhances the original Swin-UNETR by incorporating dilated attention mechanisms at each layer. We evaluated our model on L2 vertebrae segmentation within rat μCT scans, which are available to researchers upon request. Furthermore, we proposed a new segmentation loss function, AUFRC, which outperformed the standard Dice loss for this task.

The implementation of this technology holds potential for improving the efficiency and reproducibility of research outcomes in both traditional histological analysis and stereological studies. Although further validation is needed, our proposed method may ultimately contribute to enhanced patient care and facilitate the development of more targeted therapeutic interventions in the future. Overall, our model demonstrates promising results, laying the groundwork for a new class of transformer-based models with hierarchical encoders, designed to optimize performance in medical image analysis tasks.

Acknowledgements

We extend our heartfelt gratitude to the Orthopaedic Biomechanics Laboratory at Sunnybrook Health Sciences Centre for their indispensable support and contributions to our research project. We particularly wish to acknowledge the laboratory for supplying us with the vital pre-clinical rat μCT data, which significantly contributed to the success of our study. Furthermore, we express our genuine appreciation for the access they granted us to Compute Canada, which empowered us to perform the essential computations and data analysis required for our research.

Individual Contributions

- Saleh Mirzatabatabaei: Data preprocessing and augmentation for pretraining and finetuning.
- Sherry Yuan: Proposed and implemented custom dilation architectural change to Swin-UNETR model.
- Alexander Ryabchenko: Proposed and implemented AUFRC loss function and corresponding metrics.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo: “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, 2021; arXiv:2103.14030 [http://arxiv.org/abs/2103.14030].

[2] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, Daguang Xu: “Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images”, 2022; arXiv:2201.01266 [http://arxiv.org/abs/2201.01266].

[3] Yucheng Tang, Dong Yang, Wenqi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, Ali Hatamizadeh: “Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis”, 2021; arXiv:2111.14791 [http://arxiv.org/abs/2111.14791].

[4] Yixuan Wu, Kuanlun Liao, Jintai Chen, Jinhong Wang, Danny Z. Chen, Honghao Gao, Jian Wu: “D-Former: A U-shaped Dilated Transformer for 3D Medical Image Segmentation”, 2022; arXiv:2201.00462 [http://arxiv.org/abs/2201.00462].

[5] J. Nalepa et al.: "Data Augmentation via Image Registration", 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 4250-4254, doi: 10.1109/ICIP.2019.8803423.

[6] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, Leonardo Rundo: “Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation”, 2021; arXiv:2102.04525 [https://arxiv.org/abs/2102.04525].

[7] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Machine Learning in Medical Imaging. 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings Editors: Qian Wang, Yinghuan Shi, Heung-Il Suk, Kenji Suzuki. Publisher: Springer International Publishing. Book Series: Lecture Notes in Computer Science. 2017; arXiv:1706.05721 [https://arxiv.org/abs/1706.05721].

[8] N. Abraham, N.M. Khan, A novel focal Tversky loss function with improved attention U-net for lesion segmentation, in: In 2019 IEEE 16th International Symposium on Biomedical Imaging, 683–687, IEEE, 2019. arXiv:1810.07842 [https://arxiv.org/abs/1810.07842].

[9] Hamza Mahdi, Michael Hardisty, Kelly Fullerton, Kathak Vachhani, Diane Nam, Cari Whyne: “Open-source pipeline for automatic segmentation and microstructural analysis of rat knee subchondral bone”, Bone, Volume 167, 2023, 116616, ISSN 8756-3282; [https://doi.org/10.1016/j.bone.2022.116616].

[10] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science(), vol 9901. Springer, Cham. [https://doi.org/10.1007/978-3-319-46723-8_49]

[11] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation; arXiv:2105.05537 [https://arxiv.org/abs/2105.05537]

[12] Seyed-Parsa Hojjat, Michael Hardisty, Cari Whyne. (2010). Micro-computed tomography-based highly automated 3D segmentation of the rat spine for quantitative analysis of metastatic disease. Journal of neurosurgery. Spine. 13. 367-70. [https://doi.org/10.3171/2010.3.SPINE09576].

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018. [https://doi.org/10.48550/arXiv.1711.05101].

Appendix A: Figures

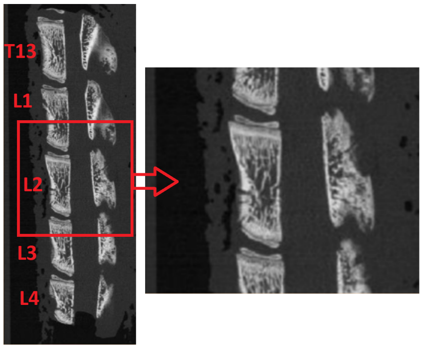


Figure 3: Location of L2 Vertebral.

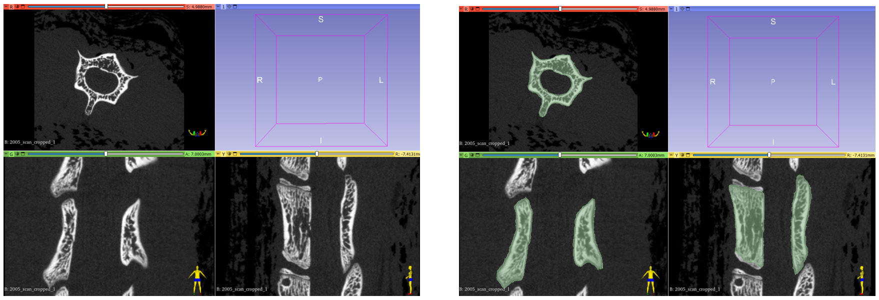


Figure 4: Example of L2-vertebrae visualized in 3D Slicer and its associated segmentation.

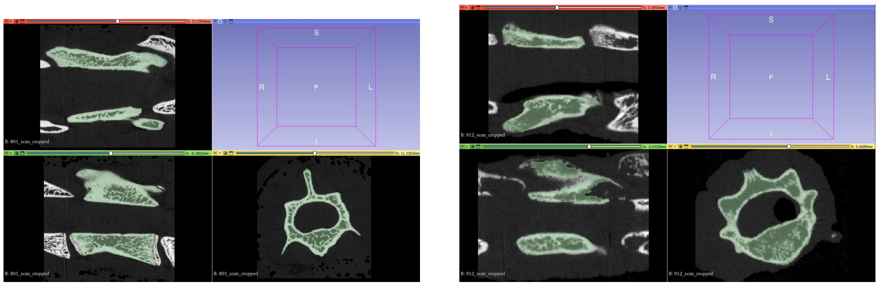


Figure 5: Qualitative visualizations of inference results. The left image is an inference on validation images. The image on the right is inference on an unlabeled sample.

	Number of Parameters	FLOPs
Dilated Swin-UNETR	63.96M	407.45G
Swin-UNETR	62.19M	396.18G

Table 2: Model configuration details.