U-Mamba2: Scaling State Space Models for Dental Anatomy Segmentation in CBCT

Zhi Qin Tan[1[0000−0002−5521−6808]], Xiatian Zhu[2[0000−0002−9284−2955]], Owen Addison[1[0000−0002−0981−687X]], and Yunpeng Li[1[0000−0003−4798−541X]]

[1] Centre for Oral, Clinical & Translational Sciences, King's College London, United Kingdom
{zhi_qin.tan,owen.addison,yunpeng.li}@kcl.ac.uk
[2] Surrey Institute for People-Centred AI, University of Surrey, United Kingdom
{xiatian.zhu,yunpeng.li}@surrey.ac.uk

Abstract. Cone-Beam Computed Tomography (CBCT) is a widely used 3D imaging technique in dentistry, providing volumetric information about the anatomical structures of jaws and teeth. Accurate segmentation of these anatomies is critical for clinical applications such as diagnosis and surgical planning, but remains time-consuming and challenging. In this paper, we present U-Mamba2, a neural network architecture designed for multi-anatomy CBCT segmentation in the context of the ToothFairy3 challenge. U-Mamba2 integrates the Mamba2 state space models into the U-Net architecture, enforcing stronger structural constraints for higher efficiency without compromising performance. In addition, we integrate interactive click prompts with cross-attention blocks, pre-train U-Mamba2 using self-supervised learning, and incorporate dental domain knowledge into the model design to address key challenges of dental anatomy segmentation in CBCT. Extensive experiments, including independent tests, demonstrate that U-Mamba2 is both effective and efficient, securing first place in both tasks of the Toothfairy3 challenge. In Task 1, U-Mamba2 achieved a mean Dice of 0.84, HD95 of 38.17 with the held-out test data, with an average inference time of 40.58s. In Task 2, U-Mamba2 achieved the mean Dice of 0.87 and HD95 of 2.15 with the held-out test data. The code is publicly available at https://github.com/zhiqin1998/UMamba2.

Keywords: U-Mamba2 · CBCT Imaging · Dental Anatomy Segmentation · Deep Learning · ToothFairy3 Challenge

1 Introduction

Cone-Beam Computed Tomography (CBCT) is a widely used imaging modality in dentistry. It provides comprehensive 3D volumetric information and excellent visualization of the orofacial region, including jaws, teeth, nerves [15]. Accurate segmentation of individual anatomical structures in CBCT images is crucial in applications such as dental diagnosis, treatment, and surgical planning [14, 20, 26]. However, manual segmentation of CBCT scans requires specialized domain

expertise and is extremely time-consuming due to their three-dimensional nature [4]. Thus, there is a strong demand for robust and efficient CBCT segmentation algorithms to improve the accuracy and efficiency of dental care and ultimately lead to better patient outcomes.

Generally, network architectures for semantic segmentation can be categorized into three: 1) Convolutional neural networks (CNN) such as U-Net [11, 22] and DeepLab [5] with translation-invariant convolutions that can effectively capture hierarchical image features and are parameter-efficient with their shared kernel weights; 2) Transformers [28] such as SETR [30] and SwinTransformer [16] that treat images as a sequence of patches instead of extracting image features hierarchically to capture the global information better; and 3) Hybrid CNN-Transformer architectures such as nnFormer [31] and SwinUNETR [24] that attempt to exploit the best of both worlds by combining their architectures.

While the hybrid architectures have improved the global feature capabilities of CNNs, transformers are highly resource-intensive due to the attention mechanism which scales quadratically with input size. This limitation reduces their suitability for healthcare applications, which often involve high-resolution 3D data and constrained computational resources in real-world settings. Recently, structured state space sequence models [9], particularly the Mamba [8] model, have emerged as an efficient and effective alternative to the transformer model. By selectively capturing relevant input features and scaling linearly with input size, Mamba outperforms transformers across multiple modalities [8, 13, 19]. U-Mamba [18] presented the first work to leverage Mamba for image segmentation, achieving superior performance and surpassing transformer-based networks in a range of medical image segmentation tasks. More recently, Dao *et al.* [7] proposed Mamba2, based on the structured state-space duality (SSD) framework, which dramatically improves speed without weakening its performance.

In this paper, we propose U-Mamba2, a hybrid CNN-SSD architecture for 3D image segmentation. U-Mamba2 extends the previous U-Mamba model [8] by leveraging the Mamba2 SSD framework that simplifies the Mamba architecture with stronger constraints imposed on the hidden space structure. Mamba2 introduced several architectural changes to enable tensor and sequence parallelism, providing a significant speedup without compromising performance. Similar to U-Mamba, U-Mamba2 can effectively extract local spatial features via CNN and capture global long-range dependencies with Mamba2. We implement interactive click prompts with cross-attention blocks and incorporate several domain knowledge to address key challenges of dental anatomy segmentation in CBCT. Our extensive experiments demonstrate the superior performance of U-Mamba2 for CBCT segmentation, outperforming previous alternatives and achieving first place for Tasks 1 and 2 of the ToothFairy3 challenge.

2 Method

This section describes our method, designed in the scope of the two tasks of the ToothFairy3 [2, 3, 17] challenges. Task 1 extends the previous ToothFairy2 chal-

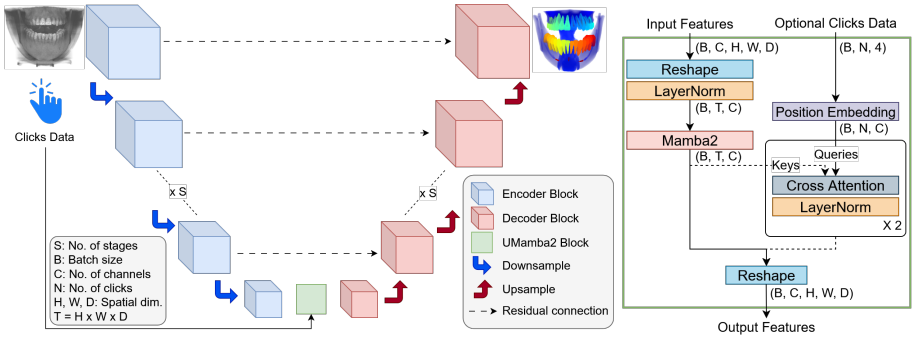


Fig. 1: (Left): Overall architecture of the U-Mamba2 model. U-Mamba2 employs the encoder-decoder framework with residual connections between each stage and the U-Mamba2 block in the bottleneck. The number of stages is configurable depending on the dataset input size. (Right): The U-Mamba2 block contains the SSD-based Mamba2 and an optional click position encoder and cross-attention blocks. The output of Mamba2 follows the solid line for tasks without interactive clicks, while it follows the dashed line when clicks are present.

lenge by adding segmentation of pulps, incisive nerves, and the lingual foramen to the existing 42 anatomy classes (*e.g.* jaws, sinuses, and 32 teeth), and includes inference time to the evaluation criteria. The dataset contains 532 CBCT scans with shapes ranging from $(170, 272, 345)$ to $(298, 512, 512)$, along with the segmentation labels of 46 anatomy classes. On the other hand, Task 2 focuses on the interactive segmentation of the inferior alveolar nerves, allowing interactive user clicks as prompts to segment the inferior alveolar nerves. Fig. 1 shows the overall structure of the U-Mamba2 model and the details of the U-Mamba2 block.

2.1 U-Mamba2: Integrating Mamba2 to U-Net

Inspired by U-Mamba [18], we propose U-Mamba2, which integrates the strengths of U-Net and Mamba2 to efficiently capture global information. As shown in Fig. 1, U-Mamba2 follows a structure similar to U-Net, with a symmetric encoder-decoder architecture that extracts image features across multiple scales. Residual connections between the encoder and decoder blocks at each stage facilitate the fusion of low-level and high-level features. As convolutional operations are inherently localized, we leverage Mamba2 to enhance the vanilla U-Net's limited capability to model global long-range dependencies in images by treating the features as long sequences. Similar to Mamba, Mamba2 scales linearly with sequence length but leverages the SSD framework to constrain the internal recurrent structure and uses matrix multiplication instead of selective scan, thereby improving efficiency through parallelism.

The encoder blocks consist of two consecutive Residual blocks [10], followed by a strided downsampling convolution, while the decoder blocks are composed of Residual blocks and transposed convolutions for upsampling. In the U-Mamba2

block, image features of shape (B, C, H, W, D) are reshaped and transposed to (B, T, C) where B denotes the batch size, C the number of channels, and H, W, D are the spatial dimensions, with $T = H \times W \times D$. Then, Layer Normalization [1] is applied to the features before they are passed to Mamba2 to capture the global contexts. The output features are then reshaped and transposed back to (B, C, H, W, D). We apply the U-Mamba2 block exclusively in the bottleneck stage, as it results in the best empirical performance for 3D computed tomography modality, consistent with Ma *et al.* [18]. Finally, Softmax is applied to the final decoder feature to produce the segmentation class probabilities, and U-Mamba2 is trained with a combination of cross-entropy loss and Dice loss.

2.2 Cross-Attention with Point Encoder

We introduce an optional interactive branch to enable the model to incorporate user-provided clicks to refine the output of U-Mamba2, improve accuracy, and support human-in-the-loop collaboration. Following the SAM2 framework [21], this branch employs a position embedding and two cross-attention blocks, as illustrated in Fig. 1. The optional clicks data contain a varying number of N clicks consisting of the X, Y, Z coordinates, and class labels. These clicks are first encoded with a learnable position embedding depending on their spatial positions and class labels. Next, the embedded click prompts and the output features of Mamba2 are fused through two-way cross-attention blocks as queries and keys, respectively. The cross-attention blocks, followed by Layer Normalization, are repeated twice to allow the model to integrate click information with the image features. The final output of the cross-attention block is then reshaped and transposed back to the original spatial dimensions.

2.3 Pre-training with Self-Supervised Learning

Recent studies [25, 27] have shown that pre-training models on large datasets with self-supervised learning (SSL) produces stronger models that can extract meaningful feature representations, leading to improved performance of downstream segmentation tasks, particularly when there is limited labeled data.

In addition to the 532 scans of ToothFairy3, we utilize the STS-3D-Tooth [29] dataset consisting of 371 unlabeled CBCT scans to pre-train U-Mamba2 with the disruptive autoencoder (DAE) [27] framework. DAE aims to reconstruct the original 3D volume after it is corrupted by several low-level perturbations. Specifically, we corrupt the input volume by randomly applying local masks, downsampling, and adding Gaussian noise to the input. The disrupted input is then passed through the U-Mamba2 to learn to reconstruct the original image with an L1 loss function. The pre-trained weights are then used to initialize U-Mamba2 (except for the weights of the optional interactive branch and the final segmentation layer) for effective downstream training.

2.4 Domain Knowledge for Dental Anatomy Segmentation

Label Smoothing of Related Anatomies. Anatomies in the orofacial region are not always distinct and often share similar shapes and properties. For instance, similar tooth types (incisor, canine, molar, premolar) between left-right counterparts, as well as the inferior alveolar and incisive nerves, exhibit close structural relationships. Therefore, to guide the model in recognizing similar classes and their spatial relationship, we introduce label smoothing for related anatomies instead of learning with hard one-hot labels. For each pixel with class k, we set the k-th class's target probability to 0.9 and distribute the remaining 0.1 evenly across the related classes. Specifically, for each voxel with a ground truth class label, k, and a set of related classes, S_r, we first initialize a zero vector, p, as the soft label, then set $p_k = 0.9$ and $p_r = \frac{0.1}{|S_r|}, \forall r \in S_r$. We apply this strategy to all anatomies with left-right counterparts, neighboring teeth, and to the inferior alveolar and incisive nerves.

Weighted Loss for Tiny Structures. ToothFairy3 introduced three additional classes corresponding to the left and right incisive nerves and the lingual foramen, which house thin, sensitive nerves in the mandible. These structures are considerably smaller than other anatomies in the dataset. We account for the volume differences by applying a class weight of 10 to these three tiny classes, so that their contribution to the overall loss is not overshadowed by larger anatomies.

Left-Right Mirroring Augmentation. The findings of the previous ToothFairy2 challenge [4, 12] showed that left-right mirroring augmentation can degrade the model's capability to reliably differentiate the left/right orientation. In dentistry, even dentists may struggle to identify a horizontally-flipped 2D image reliably without visual cues [6], due to the structural symmetry between left/right anatomies in the sagittal plane. However, we can exploit this anatomical symmetry with careful pre-processing and post-processing, enabling left-right mirroring augmentation without reducing model performance. We propose to swap the class labels of anatomies opposite to the sagittal plane whenever left-right mirroring occurs during data augmentation (*e.g.* 'Upper left canine' and 'Upper right canine'). Additionally, we also switch the predicted logits of the corresponding left/right anatomies if the image is mirrored in the left-right axis during test-time augmentation (TTA). With proper processing during training and inference, the number of possible axes combinations for mirroring augmentation is expanded from 3 to 7, substantially increasing the generalization capabilities and performance of U-Mamba2.

Post-processing. We incorporate anatomical priors of the orofacial region that voxels belonging to the same anatomy should be connected and not separated into blobs, as a post-processing step. Unlike the first-place solution [12] of the previous ToothFairy2 challenge, we perform post-processing to remove small predictions that are likely false positives based on the volume of the computed connected components [23] instead of the total volume of each class. Specifically,

Table 1: Validation set evaluation metrics. † indicates applying post-processing.

Model	Task 1					Task 2				
	Dice	HD95	Dice†	HD95†	Time	Dice	HD95	Dice†	HD95†	Time
SwinUNETR [24]	0.858	48.86	0.874	40.09	7.23	-	-	-	-	-
nnU-Net ResE [11]	0.861	45.28	0.887	32.05	**6.20**	0.901	1.98	0.905	1.71	**5.06**
U-Mamba [18]	0.865	42.06	0.896	25.88	6.98	0.903	1.65	**0.913**	1.58	5.88
U-Mamba2 (ours)	**0.873**	**41.08**	**0.908**	**21.35**	6.81	**0.905**	**1.63**	**0.913**	**1.57**	5.70

we select the threshold as the 0.5th percentile of the connected components' volume computed using the ground truth for each class. Importantly, this threshold is determined through the statistics of the ground truth rather than model predictions, ensuring that it is not model-specific. The threshold for each class is pre-computed using the entire ToothFairy3 training dataset.

3 Experiment Results

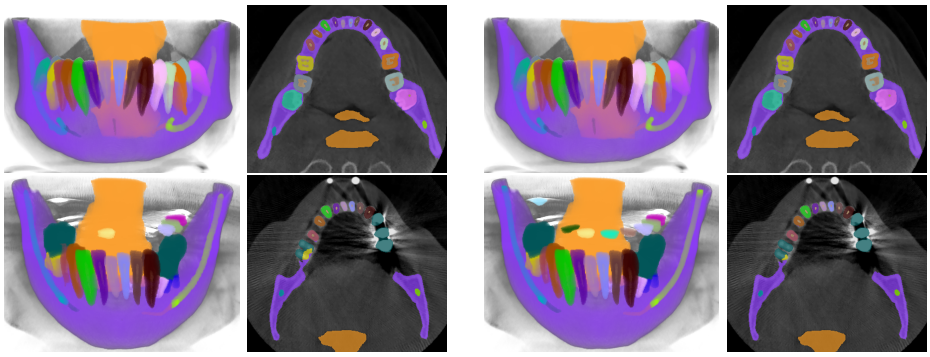
We implement U-Mamba2 with the nnU-Net [11] framework. We perform a 9:1 stratified train-validation split on the ToothFairy3 dataset to ensure the same proportion of data sources (with different fields of view and imaging machines) in the train and validation datasets. All models are pre-trained with SSL (Section 2.3) following the original training configuration of DAE [27]. Each model employs seven encoder-decoder stages, an input patch size of 128x256x256, the native voxel spacing of $0.3mm^3$ (leading to no downsampling or upsampling during model training and inference), and a batch size of 1. During training, we disable left/right mirroring augmentation for all models except U-Mamba2, while during inference, we use sliding window inference with a tile size of 0.5 and disable left/right mirroring in TTA for all models, including U-Mamba2, for fair comparison. Other hyperparameters follow the default values of nnU-Net. Model training and time computation are performed on an RTX4090 GPU. We evaluate the models with the Dice coefficient, the Hausdorff Distance at the 95th percentile (HD95), and the average inference time in seconds, where lower is better for all metrics except Dice.

3.1 Quantitative Results

Table 1 compares the proposed U-Mamba2 with nnU-Net ResE [11], U-Mamba [18] which utilizes the original Mamba layer [8], and SwinUNETR [24] on the ToothFairy3 dataset. For Task 2, we incorporate a point prompt encoder to nnU-Net ResE and U-Mamba at the bottleneck stage, similar to U-Mamba2. U-Mamba2 outperforms all benchmark models, achieving the best mean Dice score of 0.873 and 0.905 for Tasks 1 and 2, respectively. After applying post-processing, U-Mamba2 further improves to a mean Dice score of 0.908 and 0.913 for Tasks 1 and 2, respectively. U-Mamba2 delivers the best performance with an average inference time of 6.81 and 5.70 seconds per scan for the two tasks, demonstrating a slight speedup over U-Mamba.

Table 2: Ablation Study of U-Mamba2 for the validation set of Task 1. ILN indicates the metrics for the left and right incisive nerves and the lingual nerve.

Label Smoothing	Weighted Loss	L/R Mirroring	Dice	HD95	Dice (ILN)	HD95 (ILN)
×	×	×	0.867	42.36	0.617	38.41
✓	×	×	0.872	40.74	0.628	38.15
×	✓	×	0.870	41.31	0.635	37.99
×	×	✓	0.871	41.20	0.642	36.48
✓	✓	✓	**0.873**	**41.08**	**0.646**	**35.21**



(a) Ground Truth (b) Prediction

Fig. 2: Qualitative results of U-Mamba2 on the validation set of Task 1. The 3D render and a representative 2D slice are shown for: (Top) the best scoring case and (Bottom) the worst scoring case.

Furthermore, we perform an ablation study on U-Mamba2 by individually applying the dental domain knowledge introduced in Section 2.4, excluding the post-processing step (See Table 1 for post-processing results). Table 2 shows that these techniques lead to small performance improvements. In particular, the weighted loss and left/right mirroring techniques improve the mean Dice score on the three tiny structures, *i.e.* the left and right incisive nerves and the lingual nerve (ILN) from 0.617 to 0.635 and 0.642, respectively. When all three techniques are applied, U-Mamba2 achieves the best performance, with a mean Dice score of 0.873 and 0.646 for all classes and the ILN classes, respectively.

3.2 Qualitative Results

Fig. 2 visualizes the qualitative comparison between the ground truth and our model's predictions of the scans with the highest and lowest Dice score in the validation set, in the top and bottom rows, respectively. We observe that in most cases, U-Mamba2 produces precise segmentation predictions, showcasing the effectiveness of incorporating dental domain knowledge into the model design. Furthermore, we observe that U-Mamba2 can accurately localize the three

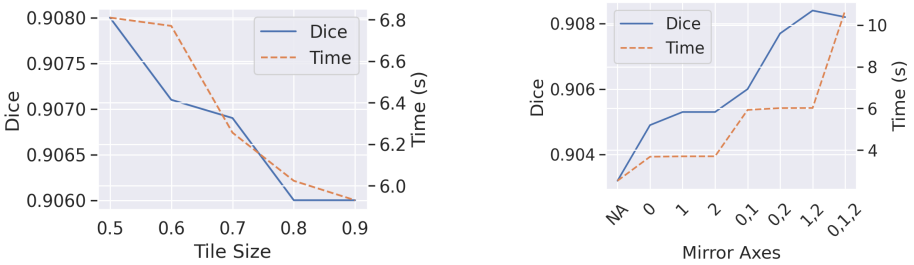


Fig. 3: (Left): Effect of the tile size on the metrics with '0,1' mirror axes in TTA. (Right): Effect of various mirror axes combinations in TTA on the metrics when tile size is set to 0.9. Axis definition: '0' is superior/inferior, '1' is anterior/posterior, and '2' is left/right.

tiny structures (ILN), producing visually acceptable segmentations. In the worst-case scenario, although the scan is imperfect due to image artifacts caused by metallic objects, false positives are primarily confined around the image edge or confusion between the actual tooth and the crown or implant, underscoring U-Mamba2's robustness under noisy conditions.

3.3 Optimizing Speed in Sliding Window Inference

As the inference time is an important metric in the ToothFairy 3 challenge, we optimize the sliding window inference parameters to improve speed without significantly deteriorating model accuracy. Specifically, we optimize the tile size parameter, where a larger value results in less border overlap during sliding window inference, and the mirror axes combinations in TTA. Fig. 3 shows the tradeoff between Dice score and inference time for different tile sizes and mirror axes combinations in TTA. By setting the tile size to 0.9, we can reduce the inference time by 12.9% with a negligible drop of only 0.002 Dice score. Moreover, Fig. 3 also demonstrates that the optimal mirror axes combination is '1,2', representing anterior/posterior and left/right, offering the best Dice score with an average inference time of only 6.02 seconds. We believe this is due to the larger spatial dimension in these axes containing more information.

3.4 Final Challenge Submission

For the final submission, we extended training to 1500 epochs using all available data with a batch size of 2 and increased the input patch size to 160x288x288. During inference, we use a sliding window inference with a tile size of 0.9 and enable mirroring in the anterior/posterior and left/right axes during TTA. The final U-Mamba2 model achieved a mean Dice of 0.84, HD95 of 38.17, with an average inference time of 40.58s, computed on the Grand Challenge platform using a T4 GPU, securing first place in Task 1 of the ToothFairy3 challenge

with a 3.1 overall ranking while obtaining first place in Task 2 with a mean Dice, HD95 and overall rank of 0.87, 2.15 and 1.66, respectively, on the hidden test set.

4 Conclusion

We presented a new architecture, U-Mamba2, designed for multi-anatomy segmentation of CBCT images in the scope of the ToothFairy3 challenge. U-Mamba2 integrates the Mamba2 SSD framework into the U-Net backbone, achieving higher efficiency without compromising performance compared to U-Mamba. By incorporating domain-specific knowledge of dental anatomy, we improved the model's performance on multi-anatomy segmentation of CBCT scans. Both the validation and independent test results demonstrate the effectiveness and efficiency of U-Mamba2, securing first place in both Tasks 1 and 2 of the ToothFairy3 challenge.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
2. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., R. Rokuss, M., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles Ballester, Vicent amd Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE Trans. Med. Imag. pp. 1–17 (Dec 2024). https://doi.org/https://doi.org/10.1109/TMI.2024.3523096
3. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volume. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 1–10. Nashville, Tennessee, USA (2025)
4. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting maxillofacial structures in cbct volumes. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 5238–5248. Nashville, TN, USA (2025)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. Eur. Conf. Comput. Vis. (ECCV). Munich, Germany (2018)
6. Chiam, S.L.: A note on digital dental radiography in forensic odontology. J. Forensic Dent. Sci. **6**(3), 197–201 (Sep 2014)
7. Dao, T., Gu, A.: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: Proc. Int. Conf. Mach. Learn. (ICML). Vienna, Austria (2024)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752 (2023)

9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). Las Vegas, NV, USA (2016)
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z
12. Isensee, F., Kirchhoff, Y., Kraemer, L., Rokuss, M., Ulrich, C., Maier-Hein, K.H.: Scaling nnu-net for cbct segmentation. In: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, MICCAI. pp. 13–20 (2024)
13. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space video models. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 87–104. Tel Aviv, Israel (2022). https://doi.org/10.1007/978-3-031-19833-5_6
14. Jaju, P.P., Jaju, S.P.: Clinical utility of dental cone-beam computed tomography: current perspectives. Clin. Cosmet. Investig. Dent. **6**, 29–43 (2014)
15. Kaasalainen, T., Ekholm, M., Siiskonen, T., Kortesniemi, M.: Dental cone beam ct: An updated review. Physica Medica **88**, 193–217 (2021). https://doi.org/https://doi.org/10.1016/j.ejmp.2021.07.007
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV) (2021)
17. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access pp. 1–12 (2024)
18. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722 (2024)
19. Nguyen, E., Goel, K., Gu, A., Downs, G.W., Shah, P., Dao, T., Baccus, S.A., Ré, C.: S4nd: modeling images and videos as multidimensional signals using state spaces. In: Proc. Adv. Neural Inform. Process. Syst. (NeurIPS). pp. 2846–2861. New Orleans, LA, USA (2022)
20. Patel, S., Durack, C., Abella, F., Shemesh, H., Roig, M., Lemberg, K.: Cone beam computed tomography in endodontics – a review. International Endodontic Journal **48**(1), 3–15 (2015). https://doi.org/https://doi.org/10.1111/iej.12270
21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024), https://arxiv.org/abs/2408.00714
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Proc. Med. Imag. Comput. Comput. Assist. Interv. (MICCAI). pp. 234–241. Munich, Germany (2015)
23. Silversmith, W.: cc3d: Connected components on multilabel 3d & 2d images (Sep 2021). https://doi.org/10.5281/zenodo.5719536
24. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 20730–20740. New Orleans, LA, USA (2022)

25. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B.A., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 20698–20708. New Orleans, LA, USA (2022)
26. Tyndall, D.A., Price, J.B., Tetradis, S., Ganz, S.D., Hildebolt, C., Scarfe, W.C.: Position statement of the american academy of oral and maxillofacial radiology on selection criteria for the use of radiology in dental implantology with emphasis on cone beam computed tomography. Oral Surg Oral Med Oral Pathol Oral Radiol **113**(6), 817–826 (2012)
27. Valanarasu, J.M.J., Tang, Y., Yang, D., Xu, Z., Zhao, C., Li, W., Patel, V.M., Landman, B.A., Xu, D., He, Y., Nath, V.: Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. In: Proc. Int. Conf. on Med. Imag. with Deep Learning. vol. 250, pp. 1553–1570 (2024)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proc. Adv. Neural Inform. Process. Syst. (NeurIPS). vol. 30. Long Beach, CA, USA (2017)
29. Wang, Y., Ye, F., Chen, Y., Wang, C., Wu, C., Xu, F., Ma, Z., Liu, Y., Zhang, Y., Cao, M., Chen, X.: A multi-modal dental dataset for semi-supervised deep learning image segmentation. Scientific Data **12**(1), 117 (Jan 2025). https://doi.org/10.1038/s41597-024-04306-9
30. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2021)
31. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. Trans. Imag. Proc. **32**, 4036–4045 (Jan 2023). https://doi.org/10.1109/TIP.2023.3293771