Tooth-Diffusion: Guided 3D CBCT Synthesis with Fine-Grained Tooth Conditioning

Said Djafar Said[1], Torkan Gholamalizadeh[2], and Mostafa Mehdipour Ghazi[1⋆]

[1] Pioneer Centre for AI, Department of Computer Science, University of Copenhagen
[2] Research and Development, 3Shape A/S, Copenhagen, Denmark

Abstract. Despite the growing importance of dental CBCT scans for diagnosis and treatment planning, generating anatomically realistic scans with fine-grained control remains a challenge in medical image synthesis. In this work, we propose a novel conditional diffusion framework for 3D dental volume generation, guided by tooth-level binary attributes that allow precise control over tooth presence and configuration. Our approach integrates wavelet-based denoising diffusion, FiLM conditioning, and masked loss functions to focus learning on relevant anatomical structures. We evaluate the model across diverse tasks, such as tooth addition, removal, and full dentition synthesis, using both paired and distributional similarity metrics. Results show strong fidelity and generalization with low FID scores, robust inpainting performance, and SSIM values above 0.91 even on unseen scans. By enabling realistic, localized modification of dentition without rescanning, this work opens opportunities for surgical planning, patient communication, and targeted data augmentation in dental AI workflows. The codes are available at: `https://github.com/djafar1/tooth-diffusion`.

Keywords: CBCT Scan Synthesis · Tooth Inpainting · 3D Generative Modeling · Conditional Diffusion · FiLM.

1 Introduction

Cone-beam computed tomography (CBCT) has become indispensable in dental and maxillofacial imaging, offering high-resolution 3D representations of dentition. However, it remains challenged by inherent limitations such as noise, metal artifacts, and a restricted field of view [1,2]. Deep learning methods have demonstrated strong potential in segmentation and reconstruction tasks, yet they often struggle with the anatomical variability of teeth, the presence of missing teeth, and the limited capacity to control or correct for structural artifacts.

Teeth segmentation from CBCT has achieved high accuracy using convolutional neural networks (CNNs), U-Net variants, and attention-based architectures [3,4,5]. However, these methods are primarily deterministic and do not support conditional generation for treatment planning, such as simulating missing

⋆ Corresponding author: ghazi@di.ku.dk

teeth, implants, bridges, or fillings. Generative models based on generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have been explored for image enhancement tasks in CBCT-to-CT and MRI-to-CBCT synthesis [6,7,8], yet prior work on generating synthetic dentition conditioned on user-specified tooth-level attributes remains scarce. To the best of our knowledge, no prior approach enables fine-grained control over individual tooth presence or absence in 3D CBCT, which is a key novelty of this study.

DDPMs have emerged as a robust framework for image synthesis due to their stability and ability to model complex distributions [9]. Recent medical imaging adaptations include CBCT-to-CT translation [7], limited-angle CBCT reconstruction [10], and medical image denoising [11]. However, these approaches have not been extended to the generation of anatomically accurate, condition-driven dental CBCTs. Such generative capabilities are clinically valuable for simulating anatomical variations, including missing or restored teeth, essential for planning personalized treatments such as implants or orthodontic interventions. Moreover, this can enhance data augmentation, address missing data scenarios, and support the training of robust models in low-resource or imbalanced datasets.

In this work, we propose a novel method to generate synthetic CBCT volumes of dentition with explicit, user-defined tooth configurations. We train a wavelet-based latent diffusion model conditioned on tooth presence, encoded via Feature-wise Linear Modulation (FiLM) embeddings [12]. By employing a masked L2 loss focused on tooth regions during training, and simulating tooth removal or addition through augmentation, the model achieves precise localization and reconstruction of dental structures. Beyond improving generative controllability, the ability to insert or remove specific teeth enables realistic pre/post-treatment simulations, supporting surgical planning, patient communication, and multidisciplinary case discussion, while also providing a targeted source of variation for augmenting datasets in tasks such as segmentation and detection.

The contributions of this paper are as follows. (1) We introduce an efficient generative framework for guided CBCT dentition synthesis, enabling explicit control over tooth presence at inference time. (2) We incorporate FiLM conditioning and a masked L2 loss to emphasize anatomically realistic reconstruction in tooth regions while suppressing background influence. (3) By simulating tooth removal and addition, we train the model to operate in two distinct modes, completion and removal, enabling scan-aware generation for clinically relevant editing (e.g., implant planning) or robust model training via data augmentation. (4) We conduct comprehensive quantitative and qualitative evaluations, including fairness across tooth positions and fidelity of reconstructed teeth, demonstrating the high realism, variability, and flexibility of the proposed model.

2 Related Work

Deep learning has improved automatic tooth segmentation, with convolutional models achieving average Dice scores above 0.9 across maxillary and mandibular scans [13]. A meta-analysis of 29 studies confirms these high accuracies and

robustness across datasets [3]. However, segmentation remains inherently limited; it does not support image generation or editing, and performance is often degraded by metal artifacts, scanner variability, and background dominance [14].

Despite growing interest in synthetic medical imaging using deep learning models, generative models have been sparsely applied to dental data. PanoGAN [15] uses a Wasserstein GAN [16] to synthesize 2D panoramic radiographs for augmentation, but it cannot capture full 3D anatomical structure. GANs also suffer from well-documented drawbacks such as mode collapse and difficulty preserving structural consistency, especially in volumetric settings.

DDPMs have emerged as a powerful alternative with greater stability and diversity in image synthesis [9]. They have been applied to CBCT denoising and CT translation, improving quality and downstream tasks like segmentation or dosimetry [7,8]. Recent variants, including DiffDenoise [11] and cycle-consistent diffusion models [10], further enhance reconstruction from sparse or artifact-prone scans. Yet, these methods treat the image volume holistically, lacking mechanisms for localized anatomical control such as tooth editing or generation.

Conditional diffusion frameworks have enabled controllable image synthesis in other domains via latent diffusion modeling (LDM) [17], cross-modal conditioning, and feature-wise transformations like FiLM. While such mechanisms support attribute-driven generation, no existing work addresses conditional 3D CBCT synthesis with anatomically precise, tooth-level guidance.

3 Methods

3.1 Guided Wavelet Diffusion Model

We employ a wavelet denoising diffusion model (WDM) [18] tailored for 3D CBCT volume synthesis. As illustrated in Fig. 1, the proposed framework follows a conditional generation paradigm, where the model is guided by binary attribute vectors representing tooth presence and a 3D CBCT scan to be edited. This conditional design enables the generation and editing of 3D scans with precise control over dental configurations.

To alleviate the high computational demands and accelerate training and inference, we replace standard Gaussian noise perturbation in pixel space with a wavelet-domain formulation. Specifically, we apply a 3D Haar wavelet transform to decompose the signal into multi-scale frequency components, and inject Gaussian noise into these components. The diffusion model then learns to iteratively denoise in the wavelet domain, operating on representations with half the original spatial resolution. This latent formulation significantly reduces memory and compute requirements while preserving semantic fidelity (e.g., global volume structure) and enhancing detail reconstruction (e.g., tooth boundaries).

3.2 Condition Embedding

Each scan is associated with a binary vector of length 32, indicating the presence or absence of individual teeth. This vector is passed through a linear layer

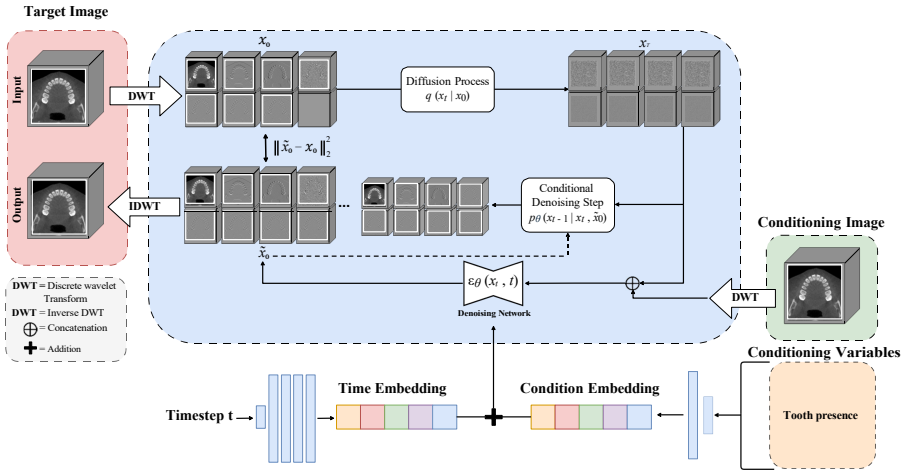


Fig. 1: Overview of the proposed framework. A guided diffusion model is used for 3D CBCT scan generation with editable tooth configurations.

to produce a learned conditioning embedding with the same dimensionality as the time embedding, which is obtained using a two-layer multilayer perceptron. These embeddings are combined within each residual block via FiLM [12], implemented as a SiLU activation followed by a linear projection, and integrated into the U-Net architecture. FiLM enables the network to modulate intermediate feature activations by applying learned, condition-dependent scaling and shifting, thereby allowing dynamic control based on the desired tooth configuration. During training, the model is conditioned on a real scan and optimized to reconstruct a version consistent with the specified tooth attributes.

3.3 Tooth Augmentation

In addition to standard training, where conditions are derived from reconstructing the input scan to match its original dental configuration, we introduce augmentation strategies to promote robustness. Specifically, we simulate two complementary scenarios by modifying the conditioning and target images: tooth addition and tooth removal. In the addition scenario, up to 50% of the teeth are randomly masked in the conditioning image, while the target remains the original, unaltered scan. The conditioning vector guides the model to plausibly reconstruct the missing structures, and the loss is computed against the original label to penalize inaccurate synthesis. In the removal scenario, up to 50% of teeth are removed from the target image, while the conditioning image retains the full dental configuration. The model learns to suppress specified regions, with supervision still applied relative to the original, unaltered label. These augmentations simulate clinically relevant use cases, such as handling missing or implanted teeth in surgical planning or restorative workflows.

To ensure that missing teeth appear realistic in the simulated training data, we avoid naive zeroing or masking of the region. Instead, we employ an image-based inpainting strategy to fill the masked tooth cavity with anatomically plausible content. Specifically, we first dilate the tooth mask to accommodate boundary uncertainty and define a broader region for removal. This expanded mask is used to set the corresponding region in the CBCT scan to missing values. Next, we apply the Manhattan (city-block) distance transform to the binary mask, indicating missing regions as zeros. The resulting distance map identifies the nearest valid voxels, whose intensity values are then propagated to fill the cavity. This approach allows the reconstructed region to reflect plausible anatomical structure, such as gradients between air and jawbone intensities near the crown and root, rather than introducing artificial holes to which the model could overfit. Finally, we apply Gaussian smoothing to the inpainted region to eliminate abrupt transitions and promote spatial coherence.

To further increase the effective training sample size, we apply a simple yet effective data augmentation by horizontally flipping the scans and their corresponding label maps (left-to-right). Given the approximate bilateral symmetry of human dentition, we adjust the tooth label values post-flip to preserve anatomical correctness. For the upper jaw (labels 1 to 16), each label is reassigned as $17-t$, and for the lower jaw (labels 17 to 32), as $49-t$, where t is the original tooth label. This transformation ensures consistency in left–right orientation and tooth identity, thereby augmenting the dataset without introducing semantic noise.

3.4 Loss Function

Given the high background-to-signal ratio in CBCT scans, we introduce a masked L2 loss to concentrate learning on tooth-bearing regions. During training, a soft spatial mask M is derived from the ground truth tooth segmentation by applying a Gaussian blur around tooth boundaries. This mask emphasizes regions near the teeth while down-weighting the less informative background. The masked L2 reconstruction loss is defined as:

$$\mathcal{L}_{\text{Masked}} = \|M \odot (x - \hat{x})\|_2^2, \tag{1}$$

where x is the ground truth scan, $\hat{x}$ is the generated output, M is the soft mask, and $\odot$ denotes element-wise multiplication. This loss penalizes discrepancies, specifically in regions affected by tooth additions or removals, encouraging anatomically faithful reconstructions. The masked loss is then combined with the primary WDM reconstruction loss after the denoising process to yield the total training objective:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{WDM}} + \lambda\, \mathcal{L}_{\text{Masked}}, \tag{2}$$

where λ is a weighting factor, empirically set to 10 in our experiments.

4 Experiments and Results

4.1 Data

We utilize a curated dataset of CBCT scans with ground truth dental segmentation[3] for our study, originally introduced in [19,20,21]. From the initial set of 150 CBCT volumes, we exclude 50 scans from a different cohort acquired at higher resolution (0.2 mm^3 voxel size) for subsequent analysis. Additionally, we discard two scans with missing segmentation maps and two segmentation volumes without corresponding scans. The resulting 98 volumes are manually relabeled according to the Universal Numbering System (tooth numbers 1–32), excluding supernumerary teeth present in two of the scans. For missing teeth annotations, we provide manual annotations where applicable. Notably, the dataset includes patients with multiple CBCT acquisitions at different treatment stages, enabling analysis of longitudinal consistency and anatomical changes.

All CBCT scans are spatially standardized to a fixed volume size of $256 \times 256 \times 256$ voxels by cropping or padding based on the tooth annotations. This ensures full dentition coverage while preserving the original voxel resolution of 0.4 mm^3, and reduces memory usage by excluding excessive background regions. Intensity values are normalized to the $[-1, 1]$ range for stable training. For each scan, binary labels are generated for individual teeth and are used both as supervision during training and for computing the masked reconstruction loss.

4.2 Experimental Setup

To scale training across multiple GPUs, we adapted the diffusion model using Distributed Data Parallel framework along with a Distributed Sampler for efficient data loading. The model was trained for 100,000 iterations with 1,000 diffusion time steps using a linear noise schedule. Optimization was performed using Adam optimizer with an initial learning rate of 1×10^{-5}, scaled linearly with the number of GPUs. A batch size of 1 per GPU was used to accommodate the memory constraints of volumetric data. During evaluation, we test the model on real CBCT scans, both with and without artificially removed/added teeth, and compare reconstructions against the corresponding ground truth volumes.

Out of the 98 available scans, we reserve 8 unique patient scans for testing, selected to represent edge cases such as complete dentition, partial dentition with only a few remaining teeth, or the presence of artifacts like brackets, braces, and mini-screws. The remaining 90 scans are used for training/validation. Although the dataset appears limited in size, each scan encompasses 32 distinct tooth states, creating a combinatorial space of present/missing patterns. Combined with data augmentation during training, this allows diverse testing scenarios in generative settings, where variability rather than sample count is critical.

We evaluate model performance using standard quality metrics for visual fidelity, including the Structural Similarity Index Measure (SSIM) for paired

[3] `https://github.com/ErdanC/Tooth-and-alveolar-bone-segmentation-from-CBCT`

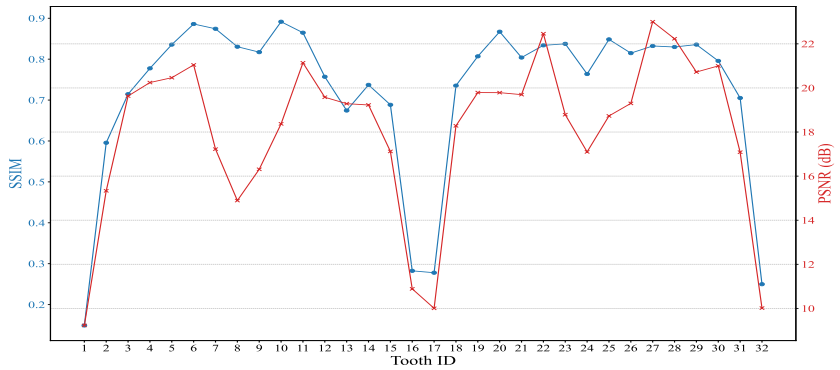


Fig. 2: Similarities between the original and reconstructed tooth when individually removed and regenerated by the model. Lower similarity is observed for wisdom teeth, likely due to their anatomical variability and data scarcity.

comparisons and the Fréchet Inception Distance (FID) for distributional comparison, both computed in 3D between real and generated volumes. In addition, we assess fairness in generation using per-tooth similarity analysis to identify potential bias in reconstruction fidelity, particularly in scans with a full dental set used for simulated tooth removal and addition scenarios.

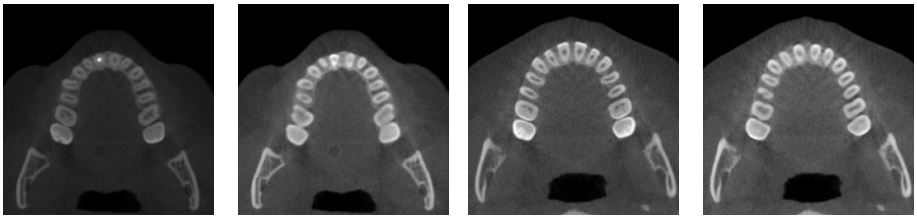
4.3 Results

Reconstruction Synthesis. We first assess the model's ability to reconstruct full CBCT scans from conditioning vectors reflecting the original dentition. Quantitative evaluation is conducted using FID between training-validation and training-test splits, providing insight into both overfitting and generalization. The FID score on the test set is notably low (40.27), indicating high-quality image generation relative to the real training samples. It also suggests that the model does not overfit or memorize the training distribution. The relatively higher FID scores for the validation set (88.81) may be attributed to the smaller number of samples (2 vs. 8) used during validation.

Tooth Addition Synthesis. To evaluate single-tooth completion, we simulate missing teeth by masking individual teeth in test scans with complete dentition. The model is then tasked with reconstructing the missing tooth based on the remaining context. We compute the SSIM and PSNR between the reconstructed and ground-truth teeth on a per-tooth basis. The average SSIM scores per tooth are visualized in Fig. 2, highlighting variation across tooth positions. As can be seen, the tooth addition results demonstrate fairly accurate synthesis across most teeth, except for the molars and wisdom teeth (i.e., tooth IDs 1, 16, 17, and 32), which are typically scarce in the training datasets and exhibit greater anatomical variability in size, shape, and orientation.

Table 1: Comparison of FID scores between test-time generated scans with removed teeth and training scans exhibiting matching tooth absence.

Missing Teeth	[1, 16]	[1, 16, 17, 32]	[16, 17, 18, 19]
FID Score	75.20	74.36	80.03



(a) Real Scan 1 (b) Synthetic Scan 1 (c) Real Scan 2 (d) Synthetic Scan 2

Fig. 3: Qualitative comparison between generated CBCT scans and their corresponding real scans with complete dentition.

Tooth Removal Synthesis. We assess the model's capacity to synthesize realistic scans with specific teeth removed. For this, we define common patterns of tooth absence and apply them to scans with full dentition in the test set. The model generates corresponding scans with these teeth removed. We then compare the generated scans to real samples from the matching tooth-absence groups using FID. Table 1 reports the FID scores for different target tooth-absence patterns. As shown in the table, the results yield low FID scores across different missing-tooth groups compared to the corresponding generated samples, demonstrating the model's ability to successfully inpaint missing teeth.

Full Dental Synthesis. As a final experiment, we evaluate the model's performance on generating a complete dentition in scans with no teeth present. This assesses the model's ability to synthesize anatomically plausible full dental structures from the conditioning vector. Fig. 3 presents qualitative results comparing the generated samples to real scans with complete dentition. The visual comparison demonstrates a strong alignment between the real and synthetic inpainted regions. Quantitative evaluation supports this observation, with an average SSIM of 0.9123 and an average PSNR of 18.35 computed over the inpainted areas, despite the model not having seen the test samples during training.

5 Conclusion

We proposed a guided diffusion framework for controllable synthesis of 3D CBCT dental scans, enabling realistic generation, addition, and removal of teeth based

on per-tooth attributes. Our approach integrated a wavelet-based denoising diffusion backbone with FiLM conditioning and masked reconstruction loss alongside tooth augmentations to guide the generative process toward anatomically plausible outputs. Experimental results demonstrated high visual fidelity and generalization performance, with low FID scores in reconstruction and inpainting tasks, and consistent SSIM values across most teeth. The model successfully handles complex cases, such as missing dentition or the presence of artifacts, and shows potential for clinically oriented simulation tasks such as visualizing treatment outcomes or testing AI models under diverse dentition patterns.

While we focused on per-tooth conditioning, broader clinical factors such as implants, crowns, and bridges remain challenging and represent promising directions for future work. Moreover, although our dataset was limited in size, each scan enabled extensive variability via combinatorial tooth presence patterns, supporting robust generative evaluation. Scaling to larger public datasets, such as ToothFairy[4], and assessing impact on downstream segmentation or detection tasks will be important future steps toward clinical deployment. This study highlights the feasibility of fine-grained, tooth-level controllable generation and provides a tool for simulation, targeted data augmentation, and the development of more customizable and interpretable generative models in dental imaging.

Acknowledgments

This project is supported by the Pioneer Centre for AI, funded by the Danish National Research Foundation (grant number P1).

Disclosure of Interests

The authors have no competing interests in the paper.

References

1. Scarfe, W.C., Farman, A.G.: What is cone-beam CT and how does it work? Dental Clinics of North America **52**(4) (2008) 707–730
2. Pauwels, R., Araki, K., Siewerdsen, J., Thongvigitmanee, S.S.: Technical aspects of dental CBCT: state of the art. Dentomaxillofacial Radiology **44**(1) (2015) 20140224
3. Sadr, S., Rokhshad, R., Daghighi, Y., Golkar, M., Tolooie Kheybari, F., Gorjinejad, F., Mataji Kojori, A., Rahimirad, P., Shobeiri, P., Mahdian, M., et al.: Deep learning for tooth identification and numbering on dental radiography: a systematic review and meta-analysis. Dentomaxillofacial Radiology **53**(1) (2024) 5–21
4. Singh, N.K., Raza, K.: Progress in deep learning-based dental and maxillofacial image analysis: A systematic review. Expert Systems with Applications **199** (2022)
5. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting maxillofacial structures in CBCT volumes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. (2025) 5238–5248

[4] `https://ditto.ing.unimore.it/toothfairy3/`

6. Choi, H., Yun, J.P., Lee, A., Han, S.S., Kim, S.W., Lee, C.: Deep learning synthesis of cone-beam computed tomography from zero echo time magnetic resonance imaging. Scientific Reports **13**(1) (2023) 6031
7. Hu, C., Cao, N., Li, X., He, Y., Zhou, H.: CBCT-to-CT synthesis using a hybrid U-Net diffusion model based on transformers and information bottleneck theory. Scientific Reports **15**(1) (2025) 10816
8. Zhang, Y., Li, L., Wang, J., Yang, X., Zhou, H., He, J., Xie, Y., Jiang, Y., Sun, W., Zhang, X., et al.: Texture-preserving diffusion model for CBCT-to-CT synthesis. Medical Image Analysis **99** (2025) 103362
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33** (2020) 6840–6851
10. Gao, Y., Pan, S., Hu, M., Xie, H., Remick, J., Chang, C.W., Roper, J., Tian, Z., Yang, X.: Limited-angle CBCT reconstruction via geometry-integrated cycle-domain denoising diffusion probabilistic models. arXiv:2506.13545 (2025)
11. Demir, B., Liu, Y., Chen, X., Chen, E.Z., Zhao, L., Mailhe, B., Chen, T., Sun, S.: DiffDenoise: self-supervised medical image denoising with conditional diffusion models. arXiv preprint arXiv:2504.00264 (2025)
12. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 32. (2018)
13. Polizzi, A., Quinzi, V., Ronsivalle, V., Venezia, P., Santonocito, S., Lo Giudice, A., Leonardi, R., Isola, G.: Tooth automatic segmentation from CBCT images: a systematic review. Clinical Oral Investigations **27**(7) (2023) 3363–3378
14. Dot, G., Chaurasia, A., Dubois, G., Savoldelli, C., Haghighat, S., Azimian, S., Taramsari, A.R., Sivaramakrishnan, G., Issa, J., Dubey, A., et al.: DentalSegmentator: robust open source deep learning-based CT and CBCT image segmentation. Journal of Dentistry **147** (2024) 105130
15. Pedersen, S., Jain, S., Chavez, M., Ladehoff, V., de Freitas, B.N., Pauwels, R.: Pano-GAN: A deep generative model for panoramic dental radiographs. Journal of Imaging **11**(2) (2025) 41
16. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, PMLR (2017) 214–223
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 10684–10695
18. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: WDM: 3D wavelet diffusion models for high-resolution medical image synthesis. In: MICCAI Workshop on Deep Generative Models, Springer (2024) 11–21
19. Cui, Z., Li, C., Wang, W.: ToothNet: automatic tooth instance segmentation and identification from cone beam CT images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6368–6377
20. Cui, Z., Zhang, B., Lian, C., Li, C., Yang, L., Wang, W., Zhu, M., Shen, D.: Hierarchical morphology-guided tooth instance segmentation from CBCT images. In: Information Processing in Medical Imaging: 27th International Conference, Virtual Event, June 28–June 30, 2021, Proceedings 27, Springer (2021) 150–162
21. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nature Communications **13**(1) (2022)