ORAL-NEXF: 3D ORAL RECONSTRUCTION WITH NEURAL X-RAY FIELD FROM PANORAMIC IMAGING

Weinan Song[1], Haoxin Zheng[1], Jiawei Yang[1], Chengwen Liang[3], and Lei He[2,1]

[1]University of California, Los Angeles
[2]Eastern Institute for Advanced Study, Ningbo, China
[3]Hangzhou Dental Hospital, Hangzhou, China

ABSTRACT

3D reconstruction of medical images from 2D images has increasingly become a challenging research topic with the advanced development of deep learning methods. Previous work in 3D reconstruction from limited (generally one or two) X-ray images mainly relies on learning from paired 2D and 3D images. In 3D oral reconstruction from panoramic imaging, the model also relies on some prior individual information, such as the dental arch curve or voxel-wise annotations, to restore the curved shape of the mandible during reconstruction. These limitations have hindered the use of single X-ray tomography in clinical applications. To address these challenges, we propose a new model that relies solely on projection data, including imaging direction and projection image, during panoramic scans to reconstruct the 3D oral structure. Our model builds on the neural radiance field by introducing multi-head prediction, dynamic sampling, and adaptive rendering, which accommodates the projection process of panoramic X-ray in dental imaging. Compared to end-to-end learning methods, our method achieves state-of-the-art performance without requiring additional supervision or prior knowledge.

***K*eywords** 3D Reconstruction · Neural Radiance Field · Dental Imaging

1 Introduction

In recent years, the development of deep learning has led to increased interest in using 2D X-ray images for 3D reconstruction in radiation imaging. Various studies, including[1][2][3][4][5][6], have employed end-to-end deep learning models to translate one or more X-ray images of the imaging target into 3D space. This technology can potentially reduce the amount of absorbed radiation during examinations, and offer an alternative imaging solution for children and elderly individuals. However, the effectiveness of such generative models heavily depends on the diversity and scale of the paired training data, where the diversity and scale of the training dataset could increase the uncertainty during inference. In the case of single-image reconstruction, the model also relies on prior knowledge of the target shape or pixel-wise annotations to reconstruct an accurate shape of the 3D object. These conditions significantly limit the widespread adoption of such learning-based tomography in clinical applications.

Compared to learning-based methods that rely on vast amounts of data, recent advances in the field of neural radiance field (NeRF) [7][8] have shown promising results in 3D object reconstruction from a single image in medical imaging. A NeRF model uses a deep neural network to approximate a field function that maps 3D positions into voxel values, and is optimized using paired poses and view images from a camera to reconstruct the 3D object. In medical imaging, the imaging equipment can provide the position of a detector and its captured image, such as an X-ray source and a digital film, making it possible to reconstruct the target 3D object without requiring additional patient data.

Several recent works demonstrate the versatility of neural radiance field (NeRF) in medical imaging applications. For instance, [9] utilizes a neural network to predict attenuation coefficients for Cone Beam Computed Tomography (CBCT) reconstruction. Similarly, [10] develops a model that disentangles the shape and depth of surface and internal anatomical structures to create a continuous representation of CT scans from a single X-ray image. In another study,

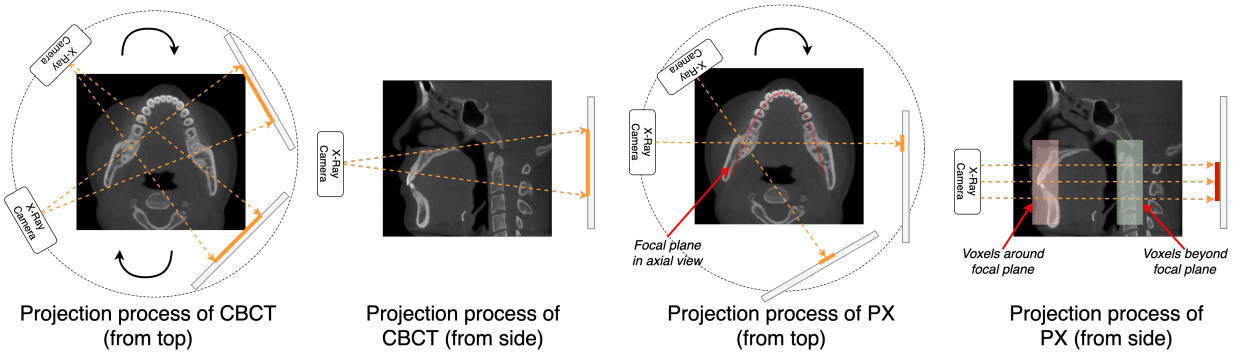

Figure 1: In CBCT, the X-ray camera rotates around a fixed center for 360° and sends cone-shaped rays to the receiver. In PX, the X-ray camera and receiver rotates around different centers during imaging for 180° to fit the curve of mandible and maxilla.

[11] applies the NeRF algorithm in 3D ultrasound reconstruction to evaluate spinal curvature measurements. Lastly, [12] presents a framework for stereo 3D reconstruction of deformable tissues from a single viewpoint in robotic surgery.

However, the existing studies have yet to investigate the application of NeRF in panoramic imaging (PX). PX has become a popular imaging technique in dental healthcare in recent decades due to its fast speed, high accuracy, and low radiation. Similar to (CBCT), which is widely used in orthodontics, the PX imaging process involves the X-ray source and sensors moving simultaneously during scanning. However, unlike CBCT, PX uses focal plane tomography to generate a projection image of the target area. The X-ray source and a moving film rotate horizontally to match the focal curve with the patient's mandible and maxilla shape, as illustrated in Figure 1. However, the limited projection information in the z dimension presents challenges for 3D reconstruction from panoramic imaging. To address this issue, we propose a framework called Oral-NeXF, based on neural field methods, for 3D oral reconstruction from panoramic imaging. We summarize our key contributions as follows:

- In contrast to previous approaches for 3D oral reconstruction, such as Oral-3D[6] and X2Teeth[3], which rely on paired data to learn an inverse projection function and prior knowledge or voxel-wise annotations to restore the curved mandible shape, Oral-NeXF does not require any paired data or individual prior knowledge.
- To overcome the limited projection information in the z dimension in PX imaging, we introduce a multi-head neural field function that predicts a beam of voxel intensities at a single time given the 2D coordinate.
- With the feature of multi-head prediction, we introduce a dynamic sampling strategy when generating point samples from radiation rays. This could encourage the model to learn a smooth intensity distribution in the 3D space.

2 Methodologies

2.1 Overview

The Oral-NeXF model is trained in a similar manner to the NeRF model, where paired projection directions and view images are provided as input, as illustrated in Fig. 2. In the training process, we generate multiple sample points along each projection ray at various sampling rates. These sampled coordinates are then used as input for a positional encoder to be transformed into high-dimensional embeddings, which are further fed into a multi-head neural field function to predict voxel intensities in 3D space. Finally, the predicted 3D points are adaptively rendered into a pixel value based on the projection direction and sampling rate. The entire model is optimized iteratively using the projection direction provided by the X-ray source and the projection image provided by the X-ray sensor. During inference, the model takes in all the coordinates around the focal plane to reconstruct the target 3D oral structure.

2.2 Neural X-ray Field

NeRF has been emerging as a promising method for reconstructing large 3D scenes from 2D images viewed at different angles and positions. Generally, a neural field function f can be taken as a function that maps a 3D coordinate into some physical attributes, e.g., color and density, which can be expressed as:

$$f_{NeRF} : (\mathbf{p}) \rightarrow (\sigma, c), \tag{1}$$

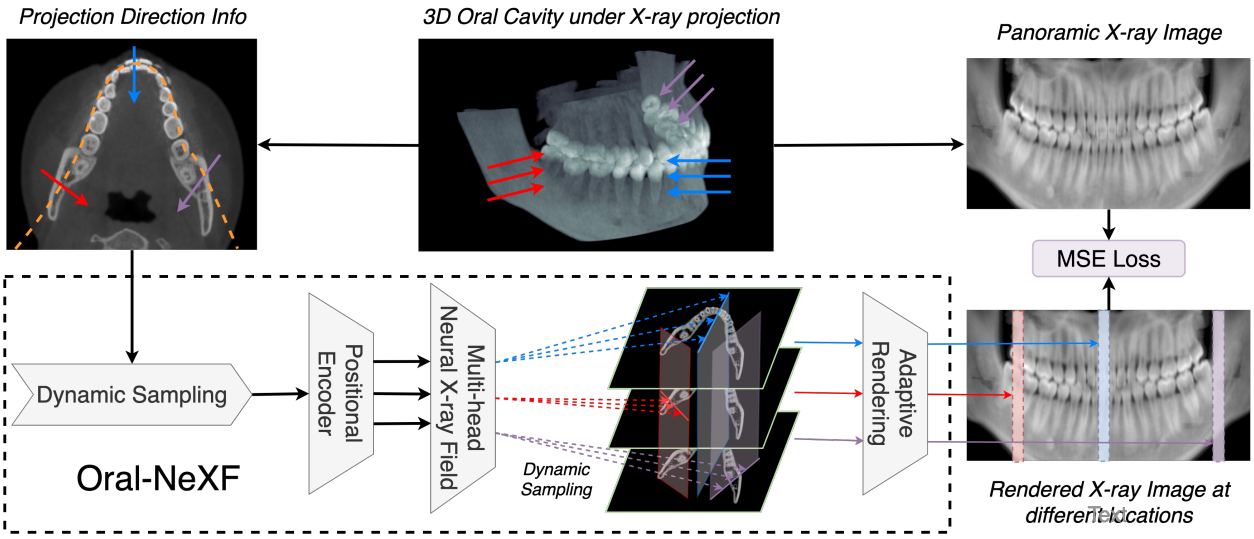

Figure 2: This image provides an overview of the Oral-NeXF model. Starting with radiation rays, we use a dynamic sampler to select projection sample points on each ray with varying sampling rates. Then, we employ a positional encoder and a multi-head neural field model to predict beams of voxel intensities in 3D space. Next, we render the projection pixels adaptively based on the sampling rate. Finally, we calculate the MSE loss between the rendered slice image and the ground-truth projection image to update the parameters of the neural field model.

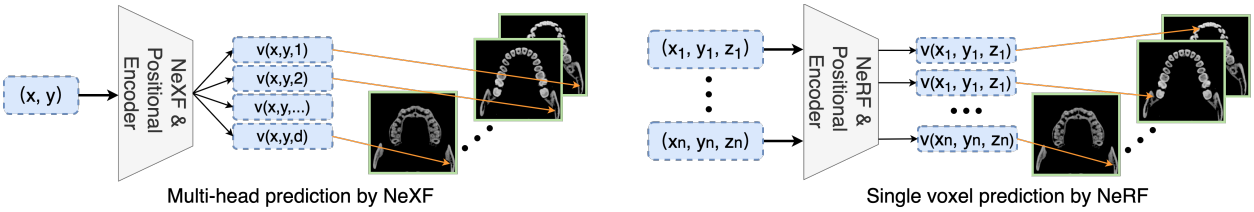

Figure 3: In this picture, we present a comparison between the predictions of NeXF and general NeRF models. The NeXF model is specifically designed for PX imaging, where the model uses a single 2D coordinate to predict intensities of voxels at the same positions in axial plane. In contrast, a general NeRF model can only generate a single voxel value given a 3D coordinate input.

where $\mathbf{p}$ is a 5D position coordinate $xyz\theta\phi$, σ is the density, and c is the color in RGB space. Given pairs of view data and render images, the model could be optimized by L2 norm between the ground-truth pixel $I_g(\mathbf{o})$ and rendered (projection) pixel $I_r(\mathbf{o})$ viewed from the position $\mathbf{o}$, which can be denoted as:

$$Loss = \mathbb{E}_{\mathbf{o}}||I_r(\mathbf{o}) - I_g(\mathbf{o})||_2 \tag{2}$$

Additionally, there is usually a positional encoder before the neural field function to exploit high-frequency variation of the target object by mapping the input coordinates into a higher embedding space. After training, the model can predict the color and intensity value at any 3D position, thus representing the 3D object implicitly by rendering from any position.

In this paper, we follow a similar framework as NeRF but change the neural field and rendering functions in different forms to accommodate PX imaging. To distinguish NeRF models, we use the term NeXF to specify the field function used in PX imaging. In general, our NeXF only predicts the voxel intensity from the spatial position due to the isotropic feature of X-ray projection. Therefore, the mapping function should be denoted as:

$$f_{NeXF} : (\mathbf{p}) \rightarrow (v_p), \tag{3}$$

where v_p is the intensity value of a voxel in the 3D space.

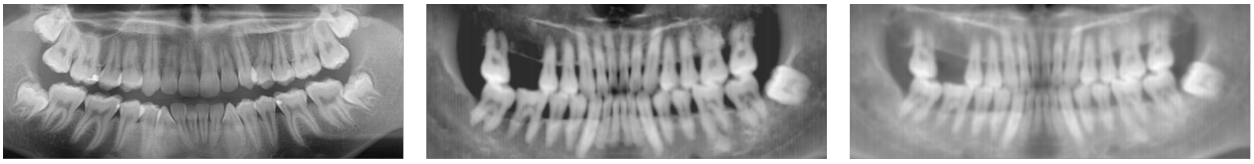

Figure 4: Comparison of different rendering methods in PX imaging. We can see that with soft rendering the generated PX image has a closer contrast with the real PX image (obtained from Internet). The real PX image looks more clear due to the high resolution of the PX machine.

2.3 Multi-head Prediction

With the feature of focal plane tomography, pixel intensities in the X-ray sensor are mainly associated with the projection space in the same axial plane, which is quite different from the imaging process of a camera or a CBCT scan. To address these limitations, we use a multi-head neural field function to predict beams of voxel intensities with the same 2D position simultaneously. Consequently, our model can output a slice of the projection image given a radiation ray and the neural field function in (3) can be further expressed as:

$$f_{NeXF} : (x, y) \rightarrow (v_{x,y,1}, v_{x,y,2}, \cdots, v_{x,y,n}), \tag{4}$$

where (x, y) is a 2D coordinate in the axial plane and n is the height of the X-ray beam. A comparison of multi-head prediction in NeXF and single-head prediction in NeRF can be seen in Fig. 3.

2.4 Soft Rendering

In this paper, we use CBCT data to simulate the PX imaging and obtain the ground truth of 3D oral structure. However, Hounsfield Units (HU) is unreliable in CBCT scans due to variations in grayscale values for different areas in the scan. This can occur even when these areas have the same density, but different relative positions within the scanned organ [13][14]. To obtain the PX image with a close distribution with real PX images, we choose to use a different rendering method following the work in [15]. Therefore, the projection value of a radiation ray that originates from the position $\mathbf{o}$ with direction $\mathbf{d}$ between t_n and t_f can be expressed as:

$$I_r(\mathbf{o}) = S \cdot \log \int_{t_n}^{t_f} e^{\frac{v(\mathbf{o}+t\cdot\mathbf{d})-C}{S}} dt \tag{5}$$

, where $v()$ denotes the voxel intensity value in 3D space, S and C are the scaling and bias factors. The selection value for C and S depend on the CBCT data used for simulation. A comparison of the proposed and conventional projection methods can be seen in Fig. 4.

2.5 Dynamic Sampling and Adaptive Rendering

To obtain a smooth intensity distribution in 3D space, we propose to use a dynamic sampling strategy to acquire points from the radiation ray. Given a projection ray from training data, we acquire points with a random sampling rate N_s. For example, the dynamic sampler utilizes different sample rates to get sample points from the red, blue, and purple radiation rays, as shown in the output of the multi-head predictor in Fig. 2. Accordingly, the rendering function in (5) could be expressed in a discrete form as:

$$I_r(\mathbf{d}) = S \cdot (\log \sum_{i}^{\lfloor N_s(t_f - t_n)\rfloor} e^{\frac{v\left(t_n + \frac{1}{N_s}\cdot i\right)-C}{S}} - \log N_s). \tag{6}$$

3 Experiments

3.1 Dataset

We obtain a dataset of 80 CBCT dental scans for our study to simulate PX imaging and obtain the ground truth for oral structure. We allocate 60 cases to train end-to-end models, while the remaining 20 cases are reserved for evaluation and optimization for NeRF-based models. We resize all CBCT scans into $288 \times 256 \times 160$ using trilinear interpolation to facilitate easy comparison.

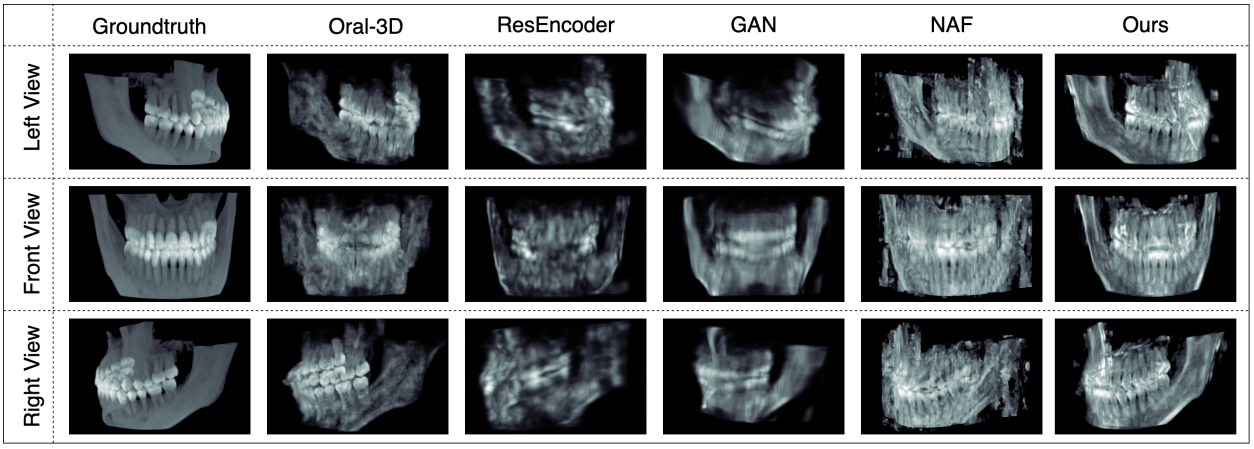

Figure 5: Comparison of 3D oral reconstruction by different methods from PX imaging. The reconstruction results are shown by maximum projection to compare density details. We could easily find that our method show the best performance with clear density density distributions and teeth boundaries.

3.2 Panoramic Imaging Simulation from CBCT

To generate projection images, we fit the focal curve using a beta function and adopt the same approach as Oral-3D. We simulate projection on 576 small curves with equal distances along the focal curve. The projection rays are simulated in a range of $\frac{\pi}{4}$ and $\frac{3\pi}{4}$ relative to each segment.

3.3 Network Architecture and Hyperparameters

We select $C = 1000$ and $S = 1200$ in Equation (6) according to the HU value of air and soft tissue. The sampling rate N_s for each radiation ray during training follows a uniform distribution in $[0.25, 1, 25]$. We use frequency embedding for the positional encoder with digital frequency at 32, and normalize the coordinates into $[-1, 1]$. For NeXF model, we use a 12-layer MLP with residual connections and set the number of heads as 160 to be consistent with CBCT data.

3.4 Training and evaluation

The model has been trained for 100k iterations with a batch size of 64 rays. The model is optimized by Adam optimizer with a learning rate starting at 0.001 decrease to 0.0001 after 20k iterations. We use structural similarity index measure (SSIM[16]), dice coefficient, and peak signal-to-noise ratio (PSNR) to evaluate the reconstruction results. We also use the averaged score proposed in [6] as the overall metric.

4 Results

4.1 Comparison of 3D reconstruction with other models

We compare Oral-NeXF with existing deep-learning-based tomography models, and present the results in Figures 5 and 1, where we observe that Oral-NeXF achieves the best performance. Oral-3D [6], ResCNN [1], and GAN [17] are trained using paired images generated from the reserved 60 cases. Specifically, GAN is trained using the same encoding-decoding network as ResEncoder and the same discriminator as Oral-3D but without any curve information. Moreover, the NAF [9] model is trained similarly to our work, but utilizes a trainable hash embedding for position encoding and a 3D attenuation coefficient predictor as the neural field function. As shown in the figures, Oral-NeXF achieves remarkable performance with clear details, without requiring prior expert knowledge or additional patient data.

4.2 Experiment analysis

Combining the results presented in Figure 5 and Table 1, we observe that Oral-NeXF achieves state-of-the-art performance. In contrast, ResEncoder and GAN can only restore the curved shape by learning from numerous paired images. Oral-3D achieves better performance in shape restoration and detail reconstruction, mainly due to prior knowledge of the dental arch shape information that enables the generator to focus on learning inverse projection. On the other

Table 1: Evaluation of 3D oral reconstruction by PSNR, SSIM, and Dice.

Method	Oral-3D	ResEncoder	GAN	NAF	**Ours**
PSNR	18.59±0.70	18.26±0.62	16.71±0.89	18.35±0.86	18.26±0.50
SSIM(%)	76.88±1.26	72.67±1.56	75.10±1.46	60.69±2.69	76.67±1.72
Dice(%)	65.94±4.24	62.52±5.56	63.96±7.03	57.20±3.94	72.09±3.63
Overall	78.60	75.49	76.93	65.93	**80.02**

Table 2: Ablation study by removing each component in Oral-NeXF. M: Multi-head Prediction, D: Dynamic Sampling, S: Soft Rendering

M	D	S	PSNR	SSIM(%)	Dice(%)	Overall
✗	✓	✓	17.12±0.86	71.28±3.38	61.03±6.07	72.64(-7.38)
✓	✗	✓	13.02±0.52	50.83±0.65	31.18±4.70	49.03(-30.99)
✓	✓	✗	15.80±0.38	58.72±0.90	53.01±3.88	63.79(-16.43)

hand, NAF fails to generate a detailed structure and contains much noise in the surroundings. As mentioned earlier, a general neural field function with 3D coordinate input and single-head prediction cannot fit PX imaging. This is also demonstrated in our ablation study.

4.3 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in Oral-NeXF. We use the letters M, D, and S to denote the experiments: 1) replacing the multi-head field function with a single-head predictor and taking in 3D coordinates as input for the positional encoder; 2) using a fixed sampling rate of $N_s = 1$ to generate sample points on projection rays; 3) changing the formula in Equation (6) to a weighted sum function that strictly follows the Beer–Lambert law by taking the voxel intensity as Hounsfield units. As shown in Table 2, the proposed dynamic sampling method plays the most crucial role in 3D reconstruction. This finding is consistent with the experiment in NeRF, where the model uses a coarse network to predict the particle density distribution for high-resolution generation. The drop in Dice and SSIM also highlights the importance of multi-head prediction and soft rendering in Oral-NeXF.

5 Conclusion

In this paper, we propose Oral-NeXF, a method for reconstructing 3D oral structures from projection information in panoramic X-ray (PX) imaging. Unlike existing deep learning models, Oral-NeXF does not require extensive patient data or dense annotations to recover the 3D object. Instead, we utilize a multi-head neural field function to predict a group of voxel values at a single time given a 2D coordinate. Furthermore, we introduce a dynamic sampling strategy and an adaptive rendering method to obtain a smooth density distribution. Extensive experiments on simulated data from 100 CBCT scans demonstrate that Oral-NeXF achieves competitive performance compared to end-to-end models, both qualitatively and quantitatively. These results demonstrate the effectiveness and efficiency of our proposed method in 3D oral reconstruction.

References

[1] Phlipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. Single-image tomography: 3d volumes from 2d cranial x-rays. In *Computer Graphics Forum*, volume 37, pages 377–388. Wiley Online Library, 2018.

[2] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10628, 2019.

[3] Yuan Liang, Weinan Song, Jiawei Yang, Liang Qiu, Kun Wang, and Lei He. X2teeth: 3d teeth reconstruction from a single panoramic radiograph. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 400–409. Springer, 2020.

[4] Yoni Kasten, Daniel Doktofsky, and Ilya Kovler. End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images. In *Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3*, pages 123–133. Springer, 2020.

[5] Hangkee Kim, Kisuk Lee, Dongchun Lee, and Nakhoon Baek. 3d reconstruction of leg bones from x-ray images using cnn-based feature analysis. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 669–672. IEEE, 2019.

[6] Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, and Lei He. Oral-3d: Reconstructing the 3d structure of oral cavity from panoramic x-ray. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 566–573, 2021.

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[8] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

[9] Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 442–452. Springer, 2022.

[10] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3843–3848. IEEE, 2022.

[11] Honggen Li, Hongbo Chen, Wenke Jing, Yuwei Li, and Rui Zheng. 3d ultrasound spine imaging with application of neural radiance field method. In *2021 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2021.

[12] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*, pages 431–441. Springer, 2022.

[13] Gwen RJ Swennen and Filip Schutyser. Three-dimensional cephalometry: spiral multi-slice vs cone-beam computed tomography. *American Journal of Orthodontics and Dentofacial Orthopedics*, 130(3):410–416, 2006.

[14] Robert T Armstrong. Acceptability of cone beam ct vs. multi-detector ct for 3d anatomic model construction. *Journal of Oral and Maxillofacial Surgery*, 64(9):37, 2006.

[15] Zhaoqiang Yun, Shuo Yang, Erliang Huang, Lei Zhao, Wei Yang, and Qianjin Feng. Automatic reconstruction method for high-contrast panoramic image from dental cone-beam ct data. *Computer methods and programs in biomedicine*, 175:205–214, 2019.

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.