# BONEMET: AN OPEN LARGE-SCALE MULTI-MODAL MURINE DATASET FOR BREAST TUMOR BONE METASTASIS DIAGNOSIS AND PROGNOSIS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Breast tumor bone metastasis (BTBM) affects women's health globally, calling for the development of effective diagnosis and prognosis solutions. While the deep learning has exhibited impressive capacities across various healthcare domains, its applicability in BTBM diseases is consistently hindered by the lack of an open, large-scale, deep learning-ready dataset. As such, we introduce the Bone Metastasis (BoneMet) dataset, the first large-scale, publicly available, high-resolution medical resource, which is derived from a well-accepted murine BTBM model. The *unique advantage* of BoneMet over existing human datasets is the repeated sequential scans per subject over the entire disease development phases. The dataset consists of over *67 terabytes* of multi-modal medical data, including 2D X-ray images, 3D CT scans, and detailed biological data (*e.g.*, medical records and bone quantitative analysis), collected from thousands of mice spanning from 2019 to 2024. Our BoneMet dataset is well-organized into six components, i.e., Rotation-X-Ray, Recon-CT, Seg-CT, Regist-CT, RoI-CT, and MiceMediRec. We further show that BoneMet can be readily adopted to build versatile, large-scale AI models for managing BTBM diseases in terms of diagnosis using 2D or 3D images, prognosis of bone deterioration, and sparse-angle 3D reconstruction for safe long-term disease monitoring. Our preliminary results demonstrate that BoneMet has the potentials to jump-start the development and fine-tuning of AI-driven solutions prior to their applications to human patients. To facilitate its easy access and wide dissemination, we have created the BoneMet package, providing three APIs that enable researchers to (i) flexibly process and download the BoneMet data filtered by specific time frames; and (ii) develop and train large-scale AI models for precise BTBM diagnosis and prognosis. The BoneMet dataset is officially available on Hugging Face Datasets at https://huggingface.co/datasets/BoneMet/BoneMet. The BoneMet package is available on the Python Package Index (PyPI) at https://pypi.org/project/BoneMet. Code and tutorials are available at https://github.com/BoneMet/BoneMet.

## 1 INTRODUCTION

Breast cancer stands as one of the most prevalent non-skin cancers affecting women globally (6). Although the survival rate for breast cancer patients has been greatly improved due to the adoption of cancer screening, early diagnosis, and more effective cancer treatments, the metastasis of breast cancer to other organs including bone dramatically reduces the five-year survival rate and worsens the patient suffering including severe pain, impaired mobility, and elevated risk of fatal fractures (10). Once invading the bone, breast cancers activate bone resorption, which in turn drives cancer growth. The "vicious cycle" between tumor and bone leads to the rapid decline of bone structural integrity (11). The conventional approaches for managing breast tumor bone metastasis (BTBM) diseases utilize various medical imaging modalities including i) 2D X-ray images (12; 24), ii) 3D computed tomography (CT) scans (25; 3), and iii) combined localized metabolic imaging and CT scans (7; 36). However, these imaging based diagnosis and prognosis approaches suffer from several limitations, such as low sensitivity to small lesions in early-stage bone metastases (9), health risks of CT due to the exposure to high ionizing radiation (5), and interpretation variability limited by radiologist's experience and expertise (26). The lack of sensitivity and specificity of the imaging tools and algorithms leads to the

current clinical guidelines against active monitoring of BTBM because the harms of false positive diagnosis and radiation from longitudinal imaging outweigh potential benefits (14). This poses a classic catch 22 dilemma. With clinicians being hesitant to image, we cannot feed and train AI models for better diagnosis. Conversely, without effective early detection tools, we miss the critical window of treating BTBM and fail to improve patient outcomes in a meaningful way.

Deep learning-based methods have shown significant improvements in identifying subtle patterns and features that are likely missed by human eyes, in particularly when analyzing large volumes of longitudinal data, reducing the scanning duration (radiation exposure) with efficient 3D CT reconstruction, and providing consistent and objective analysis and more comprehensive insights into patient health (21; 20; 34; 28). For BTBM diagnosis and prognosis applications, previous studies (23; 37; 17) have attempted to leverage deep learning models, while their performances appear to be hindered with low resolution images, inadequate diagnosis/prognosis labels (16), limited dataset sizes (1; 38; 13), and lack of diverse modalities (18). Hence, although the utilization of AI models to manage BTBM diseases offers a promising direction, the scarcity of large-scale, high-quality, relevant datasets—containing high-resolution X-ray and CT images along with detailed medical records—greatly impedes the development and application of large-scale AI models.

To further unleash the power of deep learning in BTBM diagnosis and prognosis, we first attack the major obstacle—the scarcity of high-resolution image datasets of early-stage BTBM with sufficient diagnostic and prognostic labeling—with our open-sourced large-scale BoneMet datasets. We leverage the large preclinical images acquired over the past five years from well-established mouse BTBM models (30). Mouse models, despite limitations, are valuable tools in elucidating disease mechanisms, identifying diagnostic biomarkers, and testing treatments (8). The small-sized mouse skeleton allows quick three-dimensional CT scans (< 4 min per scan) at high resolution (7-10 micron per pixel) and low radiation exposure (<0.4 Gy) for multiple time points over the span of breast cancer metastasis (30). Additional benefits include i) the complete control/documentation of the age of the experimental mice, the bone site of cancer cell metastasis, and the type and number of breast cancer cells being introduced in each mouse, ii) the application of various treatments to mimic human patient conditions such as chemo and radiation therapy, as well as exercise regimen (31). The rigorous experimental design contains age-matched non-tumor and placebo-treated controls, and thus the acquired image dataset reflects the complexity of human breast cancer progression with superior image quality and labeling over available patient data. We further demonstrate the utility of our BoneMet datasets in supporting the development of imaging biomarkers and AI applications, such as multimodal large vision models (MLVM) with temporal and spatial alignment, for the early diagnosis and prognosis of breast tumor bone metastasis as well as improved predictive accuracy of the models used in clinical settings.

Specifically, the dataset is collected from thousands of mice spanning from 2019 to 2024 (*i.e.*, 5 years), with its total size over *50 terabytes*. Each mouse undergoes 4 to 5 weekly sequential skeletal scans, including scans of the tibiae, femurs, and vertebrae. Each session produces 260 high-resolution, multi-angle 2D X-ray images at 0.8° intervals. From these X-ray images, 3D reconstructed CT images, segmented and registered CT scans, and RoI-cropped CT images of the tibiae are obtained using processing tools such as commercial software and our developed APIs. In addition, detailed biological data is recorded for each mouse, capturing critical medical details such as age, body weight, sex, metastatic tumor growth status, bone structural and mechanical properties, and other relevant data at each time point. To facilitate its effective use, this dataset is well-organized into six key components: Rotational X-Ray Imagery **(Rotation-X-Ray)**, Reconstructed CT Imagery **(Recon-CT)**, Segmented CT Imagery **(Seg-CT)**, Registered CT Imagery **(Regist-CT)**, Region of Interest CT Imagery **(RoI-CT)**, and Mice Medical Records **(MiceMediRec)**. Figure 1 provides an overview of our BoneMet dataset, showcasing examples of its six components, the processing tools used to acquire them, and the detailed data collection procedures. Besides, we have developed the BoneMet package, which includes three types of APIs for CT image segmentation, CT image registration, and RoI-based CT image cropping, respectively. These tools have been released on the Python Package Index (PyPI), aimed at assisting researchers and practitioners by providing efficient, standardized methods for data processing. These APIs enhance reproducibility, streamline the processing of large datasets, offer user-friendly interfaces, and ensure consistent data handling.

To the best of our knowledge, this dataset is the first of its kind, offering large-scale, high-resolution, and multi-modal medical data specifically targeting the management of BTBM diseases. It can support both supervised BTBM diagnosis and prognosis, thanks to its ground-truth labels, and self-

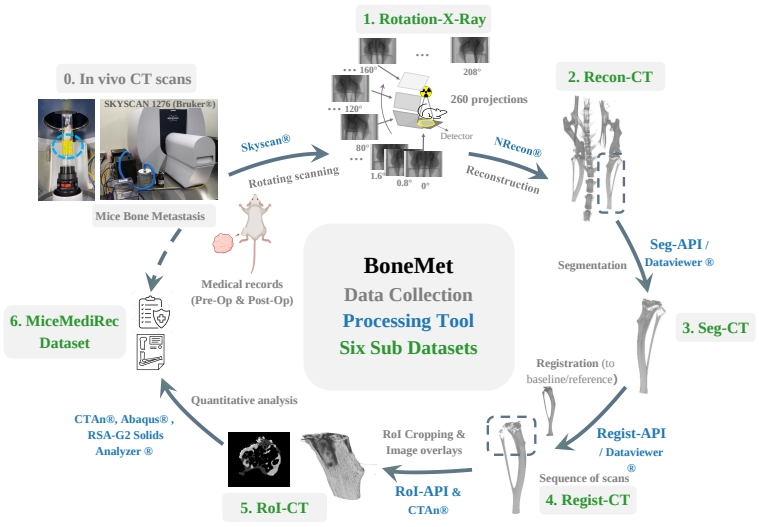



Figure 1: The data collection, processing tools, and six components for our BoneMet dataset.

supervised pre-training, due to its vast amount of data samples. Our experimental results validate the comprehensive applicability of our BoneMet dataset.

## 2 Our BoneMet Dataset

Our BoneMet dataset is an open, large-scale, and multi-modal resource specifically designed for BTBM diseases. It offers high-resolution multi-modal medical data, including 2D X-ray images, 3D CT scans, and medical records and quantitative analysis. This section details the collection and preparation process of BoneMet dataset, its contents, and potential applications.

### 2.1 Data Collection and Preparation

This dataset has spanned five years, from 2019 to 2024, and includes data over 1000 either C57BL/6J or Balb/C mice with the age varying from young adult (10-12 weeks) to aged (>70 weeks) mice. These mice were divided randomly assigned into tumor and non-tumor groups, which further divided to receive individual or combined cancer interventions such as exercise, radiation therapy, chemotherapy and other experimental drugs (30). Breast cancer cells (1000 Py8119 cells or 3000 4T1 cells) were directly injected into the proximal end of mouse tibiae of the tumor groups. The tumor and non-tumor carrying mice undergoing interventions or placebo treatments were then longitudinally monitored by weekly microCT scans over a period of 3-5 weeks. The overview of data collection and preparation procedures is illustrated in Figure 1, and the details are presented as below, which produce the BoneMet dataset into six well-organized components.

*First,* the **Rotational X-Ray Imagery (Rotation-X-Ray)** is captured weekly by the SkyScan® 1276 scanner (Bruker, Kontich, Belgium), starting from the beginning of the breast tumor inoculation to the end of the sacrifice of the mice for 3 to 5 weeks. The camera revolved around the anesthetized animal's body, covering a total rotation of 208°, from the anterior (front view) to the lateral and posterior (back view) aspects of the hind limbs with a step of 0.8 degrees, and there are 260 X-ray projections in total for each scan. The Rotation-X-Ray dataset collected in each week are organized for temporal alignment. The total scanning process for each mouse takes around 10 minutes.

*Second,* these 260 X-ray images taken from variable angles were reconstructed by NRecon® with a voxel size of 7 to 10.6 µm as the **Reconstructed CT Imagery (Recon-CT)**. The 3D reconstructed CT is acquired based on the traditional filtered backprojection (FBP), which uses a 1D filter on the projection data before backprojecting (2D or 3D) the data onto the image space. Then, the 3D bone reconstructed CT were optimized to get the high quality reconstructed CT images. The reconstruction process takes over 30 minutes per mouse.

*Third,* among the Recon-CT imagery, both left and right tibiae were segmented from the knee by either our Seg-API or Dataviewer® software as the **Segmented CT Imagery (Seg-CT)**. A global threshold

value (75/255) was applied and the processed images were found to agree well with gray-scaled images. The segmentation of both left and right tibiae takes around 10 to 15 minutes per mouse.

*Fourth,* after segmentation, each tibia was registered to their reference (vertically aligned tibia) by our Regist-API or Dataviewer® software as the **Registered CT Imagery (Regist-CT)**. For the first scan (week 0) of the time sequence, each segmented tibia is aligned to a reference tibia with its long axis vertical and its anterior-posterior and medial-lateral axes orthogonally positioned. Subsequent scans of the segmented tibia captured at later time points are aligned to themselves baseline (week 0). The registration process of each tibia to their references takes about 15 minutes.

*Fifth,* the region of interest of tibiae (proximal end where the metastatic tumors are located) overlaid to their baseline was selected by our RoI-API and the fibular were cropped manually by CTAn® as the **Region of Interest CT Imagery (RoI-CT)**. The overlapped composite RoI-CT images were generated with different pixel values assigned in each pixel based on various criterion: the pixel with bone (above threshold value of 75) both in baseline and sequential CTs were assigned to the value of 180 (light gray), the non-bone pixel (below threshold) at week 0 later became bone (above threshold) is assigned to 240 (white), the bone pixels (above threshold) at week 0 later became non-bone (below threshold) is set to 60 (dark gray), the pixel without bone in both time points were given the value to 0 (black). After the generation of CT composite, the RoI section was selected with the proximal tibia-fibula junction as the landmark, and the fibula was manually cropped by CTAn®. This process takes approximately 20 minutes for each registered tibia.

*Sixth,* the quantitative analysis of the RoI-CT imagery by CTAn®, Abaqus®, and RSA-G2 Solids Analyzer®, and the Pre-Op (Pre-Operation) and Post-Op (Post-Operation) medical records of the mice with bone metastasis were combined as the **Mice Medical Records (MiceMediRec)**. The imaging diagnosis parameters include records of tumor growth status (if available), bone structural properties, and mechanical properties. Tumor burden and growth are quantified via IVIS imaging of luminescence-tagged tumor cells, with the average radiant efficiency measured from the tibia reported. Bone structural properties, such as bone lesion occurrence, total bone volume changes, cortical bone parameters (resorption, area, and thickness), and trabecular bone parameters (bone volume fraction, number, thickness, and spacing), as well as bone mechanical properties (polar moment of inertia (pMOI)) are measured by CTAn® based on the CT in the RoI-CT dataset. For some samples, additional mechanical properties such as bone stiffness, yield strength, ultimate strength, and work to fracture are simulated or measured by finite element analysis or 3-point bending tests. This process takes over 30 minutes for each mouse.

Due to the page limit, further details on data collection and preparation processes, including breast tumor tibial inoculation and medical image processing, are provided in Section A of the supplementary materials.

### 2.2 Details of BoneMet Dataset

The total size of our BoneMet dataset is 67.87 TB, which includes six components, detailed as below. The former five components are all visual data, stored in PNG format, while the last one is numerical data, stored in CSV format.

**Rotation-X-Ray.** The Rotational X-Ray Imagery consists of 651,300 X-ray images of subjects with tumors and 676,000 X-ray images of subjects without tumors. Each image has a resolution of 4,032x4,032x1 pixels and a spatial resolution of 0.8°, captured at the hindlimb. This dataset has been aligned both spatially and temporally with the temporal resolution of 1 week, and it offers 2D X-ray images taken from multiple angles, from anterior (front) to lateral (side) and posterior (back) views, providing a comprehensive examination of the subject. The total size of this imagery is 20.93 TB. The left part of Figure 2 shows examples of the 2D X-ray images.

**Recon-CT.** This Recon-CT dataset comprises 3D CT scans reconstructed from Rotation X-Ray imagery and therefore it's also temporally aligned with the temporal resolution of 1 week. The reconstruction process is illustrated in the right part of Figure 2. These slices capture cross-sectional views of the tibia, femur, and spine, with the dimensions varying according to the specific regions of interest (RoI) identified during the micro-CT reconstruction process. This component includes 2,505 CT scans of subjects with tumors and 2,600 CT scans of subjects without tumors. Each CT scan is composed of 2,685 2D slices with an image resolution of approximately 2,588x2,428x1 pixels, with an example shown in the left image of Figure 3. The total size of this dataset is 45.23 TB.

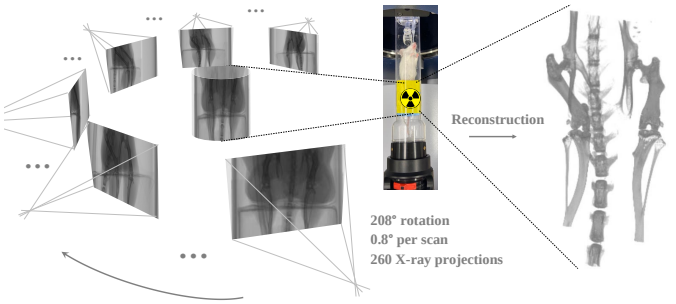



Figure 2: Examples of our Rotation-X-Ray and Recon-CT imagery. **Left:** 2D X-ray images are taken from 260 variable angles with 0.8° intervals of the hindlimb of a mouse; and **Right:** The paired 3D CT scan is reconstructed from these 2D X-ray images.

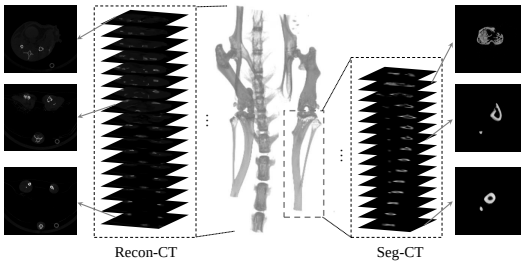



Figure 3: Examples of 3D CT scans from the Recon-CT and the Seg-CT imagery. **Left:** 3D CT scans of hindlimbs in the Recon-CT imagery; and **Right:** 3D CT scans of segmented tibiae in the Seg-CT imagery. Notably, 3D CT scans are composed of 2D cross-sectional slices.

**Seg-CT.** These 3D CTs of tibiae are isolated from the 3D CT scans of hindlimb in the Recon-CT imagery, as illustrated on the right side of Figure 3. This component includes 3,005 segmented CT scans of subjects with tumors and 7,205 segmented CT scans of subjects without tumors. Each scan is composed of approximately 1,700±200 2D slices with an image resolution of approximately 700±50x900±80x1 pixels. The size of this dataset is 1.53 TB.

**Regist-CT.** This imagery includes registered 3D CT scans of tibiae taken at various time points and from different animals, aligned to a reference, as shown in Figure 1 at Section A of supplementary materials. This component includes 3,005 registered CT scans of subjects with tumors and 7,205 registered CT scans of subjects without tumors. Each scan is composed of 1,538 2D slices with an image resolution of 509x539x1 pixels. The size of this dataset is 0.18 TB.

**RoI-CT.** This imagery focuses on the proximal end sections of the registered tibiae, where the effects of metastasis are most pronounced, as shown in Figure 4. The RoI-CT imagery comprises 300 2D slices below the proximal tibia-fibula junction, with overlaid registered CT scans aligned to their baseline (week 0). In each 2D slice, light gray represents the reserved bone in the sequential scans, white indicates bone formation where non-bone pixels at week 0 later became bone, and dark gray signifies bone resorption where bone pixels at week 0 later became non-bone. This component includes 3,005 CT scans of the proximal end sections of registered tibiae with tumors and 7,205 CT scans of that without tumors. Each 2D slice has the image resolution of 509x539x1 pixels. The size of this dataset is 8.00 GB.

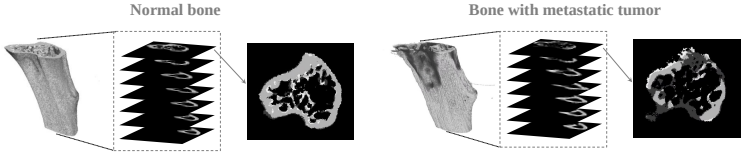



Figure 4: Examples of 3D CT scans from the RoI-CT imagery, including the proximal end of tibiae without (**Left**) and with (**Right**) metastatic breast tumors.

Table 1: Overview of our BoneMet dataset details.

| Datasets | Size | Number of Images/Records (Positive: mouse receiving tumor injection; Negative: normal non-tumor mice. The label is assigned on the animal level, and the timing of overt bone lesion varies due to age and treatment.) | Data Format | Image Size | Anatomical Region | Temporal Resolution | Data Content |
|---|---|---|---|---|---|---|---|
| Rotation-X-Ray | 20.9 TB | Positive: 651,300<br>Negative: 676,000 | 2D X-ray | 4,032 x 2,688 x 1 | Hindlimb (0.8° intervals) | 1 week | X-ray images from the anterior (front view) to the lateral and posterior (back view) |
| Recon-CT | 45.2 TB | Positive: 2,505<br>Negative: 2,600 | 3D CT<br>(Each with 2,685 slices) | 2,588 (±80) x<br>2,428 (±100) x 1 | Hindlimb | 1 week | Stacked 2D cross-sections of CT spatially and temporally paired with Rotation-X-Ray |
| Seg-CT | 1.5 TB | Positive: 3,005<br>Negative: 7,205 | 3D CT<br>(Each with 1,700 (±80) slices) | 700 (±50) x<br>900 (±80) x 1 | Tibiae | 1 week | The segmented tibiae from Recon-CT |
| Regist-CT | 0.2 TB | Positive: 3,005<br>Negative: 7,205 | 3D CT<br>(Each with 1,538 slices) | 509 x 539 x 1 | Tibiae | 1 week | The registered tibiae dataset aligned to a vertically aligned reference |
| RoI-CT | 8.0 GB | Positive: 3,005<br>Negative: 7,205 | 3D CT<br>(Each with 300 slices) | 509 x 539 x 1 | Proximal end of tibiae | 1 week | The selected CT from region of interest (RoI) of bone metastasis |
| MiceMediRec | 9.4 MB | Positive: 501<br>Negative: 520 | Text | NA | NA | 1 week | Medical record and quantitative analysis results of 3D CT images, simulation and mechanical testing |

**MiceMediRec.** The Mice Medical Records includes the number of mice of 501 with tumor and of 520 without tumor and results from detailed medical records such as experiment date, animal ID, age, body weight, mouse strain (or genotype), sex, and specific metastatic tumor sites, and quantitative analyses of bone from CTs, FE simulations, and mechanical testing, offering a comprehensive overview of the animals, bones, and their disease conditions. More details can be found in Supplementary Materials, Section A, Table 2S. The size of this dataset is 9.44 MB.

Table 1 lists more detailed and comprehensive information for our BoneMet dataset and its six components.

### 2.3 Potential Applications

There are many applications can be supported by BoneMet database. Due to page limits, we only validate some critical applications in Section 3, such as i) 2D spatial-temporal X-ray imagery-based BTBM diagnosis; ii) 3D CT imagery-based BTBM diagnosis; and iii) 3D multi-modal RoI-CT imagery-based BTBM prognosis; and iv) Sparse-angle 3D CT reconstruction with one real 2D X-ray images. Besides, more supportive applications, such as 3D CT auto-segmentation, generalist biomedical AI diagnosis, among others, are listed in Supplementary Materials, Section A, Table 1S to exhibit its broader applicability.

## 3 Experiments and Results

We conducted experiments on our BoneMet dataset via developing various deep learning solutions to exhibit its applicability and efficiency to manage BTBM disease. Currently, the positive and negative labels are assigned at the animal level, rather than individual time points and individual X-ray images. A positive label of a mouse indicates that a metastatic bone lesion occurs in the subject between week 0 and week 5. There are animal-to-animal variations at the times of bone lesion initiated and the speed of lesion growth.

### 3.1 Hyperparameters and Dataset Splitting

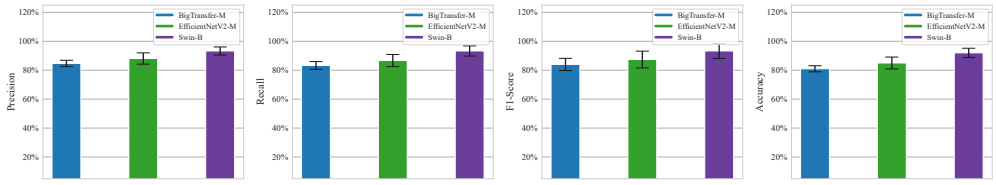
The hyperparameters used in our Breast Tumor Bone Metastasis (BTBM) diagnosis experiments follow the common practice of supervised ViT training. The key hyperparameters are detailed in the Table 2. For the experiments presented in our manuscript, we employed a train/test split with a ratio of 8:2. Specifically, 80% of the mice were used for training the model, and the remaining 20% were reserved for testing. Within the training dataset, we implemented a 5-fold cross-validation strategy. This process was

Table 2: Training Configuration

| Config | Value |
|---|---|
| Optimizer | AdamW |
| Base learning rate | 1e-3 |
| Weight decay | 0.05 |
| Optimizer momentum | $\beta_1, \beta_2 = 0.9, 0.999$ |
| Learning rate schedule | Cosine decay |
| Warmup epochs | 5 |
| Training epochs | 100 |
| Augmentation | RandAug (9, 0.5) |

Table 3: The BTBM diagnosis using ViT with and without STA on the Rotation-X-Ray imagery

| Methods | Training | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Accuracy | Precision | Recall | F1-Score | Accuracy |
| ViT (w/o STA) | 99.6 | 99.7 | 99.7 | 99.5 | 92.3 | 80.5 | 86.0 | 79.1 |
| ViT (w/ STA) | 95.4 | 96.9 | 96.1 | 95.0 | 92.1 | 90.6 | 91.3 | 89.0 |




(a) Precision (b) Recall (c) F1-Score (d) Accuracy

Figure 5: The diagnosis of BTBM using 3D CT scans from the Regist-CT imagery.

repeated five times, with each subset serving as the validation set once, so the validation data were never seen by the model during the training phases, ensuring that there was no data leakage. In our trials, the data split was conducted at the mice level, so the entire sets of images from individual mice were allocated either to the training or testing datasets. This was to ensure no overlap of data from the same mice between the training and testing datasets, preventing any potential data leakage. All our experiments were conducted on a high-performance workstation equipped with an NVIDIA RTX A6000 GPU, which has 48GB of VRAM.

### 3.2 Applicability of Our Rotation-X-Ray Dataset for BTBM Diagnosis

We conduct experiments to demonstrate the applicability of our Rotation-X-Ray dataset to manage BTBM disease. Here, we employ two Vision Transformers (ViT) variants for the diagnosis of BTBM using Rotation-X-Ray imagery: a simple ViT variant Swin-Base (22), and a spatial-temporal ViT variant MMST-ViT (19). Notably, the 2D X-ray images within the Rotation-X-Ray imagery have been carefully spatially and temporally aligned. For clarity, we denote MMST-ViT as ViT (w/ STA), indicating its ability to leverage spatial-temporal alignment (STA), and Swin-Base, as ViT (w/o STA), indicating not utilizing STA. Table 3 presents the performance outcomes measured by the metrics of Precision, Recall, F1-Score, and Accuracy. We have two observations. First, both ViT (w/o STA) and ViT (w/ STA) can achieve decent diagnosis performance results, with the overall accuracies of 79.1% and 89.0%, respectively. This suggests that our dataset is adaptable to various ViT variants, from simple to more complex architectures, for effective BTBM diagnosis. Second, we observe a significant training-test accuracy gap of 20.4% with ViT (w/o STA), indicating a pronounced overfitting issue inherent to the ViT architecture. In contrast, the incorporation of STA in the Rotation-X-Ray imagery substantially alleviates this issue, as demonstrated by the notably smaller training-test accuracy gap of 6.0% achieved by ViT (w/ STA). This highlights the effectiveness of STA in mitigating overfitting and enhancing model generalizability on our Rotation-X-Ray dataset.

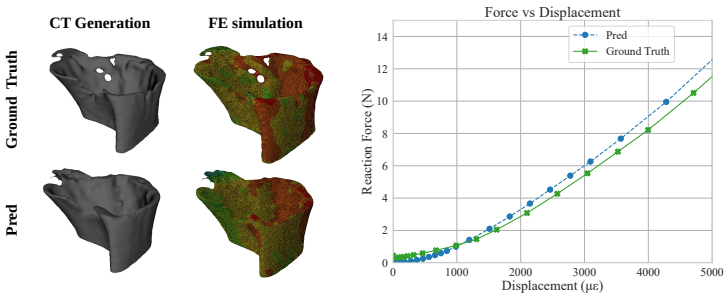
### 3.3 Applicability of 3D CT Scans for BTBM Diagnosis across Various Model Architectures

Here, we conduct experiments by utilizing 3D CT scans in the Regist-CT imagery for BTBM diagnosis. We utilize two CNN-based model architectures—BigTransfer (BiT-M) (15) and EfficientNetV2-M (27)—along with one ViT-based model architecture, *i.e.*, Swin-B, as the backbone networks. These models extract slice-level features from each 2D slice within the CT volume and subsequently aggregate these features into volume-level representations using a max-pooling layer.

Figures 5a, 5b, 5c, and 5d present the performance results. We observe BiT, EfficientNetV2, and Swin-B exhibit commendable performance, achieving overall accuracies of 81.0%, 85.0%, and 92.0%, respectively. These results validate that our BoneMet dataset is compatible with both CNN-based and ViT-based model architectures, demonstrating its broad applicability. Furthermore, the Swin model consistently outperforms its two CNN-based counterparts. For instance, in terms of the F1-Score, it exceeds BiT and EfficientNetV2 by 9.3% and 5.9%, respectively. This superior performance is

Table 4: Evaluations of the quality of 3D CT prediction using three metrics

| Methods | PSNR (↑) | SSIM (↑) | LPIPS (↓) |
|---|---|---|---|
| 3D-GAN | 21.9 | 0.767 | 0.098 |
| T-VAE | 23.4 | 0.817 | 0.078 |
| ST-VAE | 35.5 | 0.860 | 0.041 |




(a) CT generation and FE simulation (b) Reaction force prediction

Figure 6: Evaluations of prognostic assessment of bone mechanical properties: **(a)** comparisons between ST-VAE-generated 3D CT scan and its ground truth, along with the corresponding finite element (FE) analysis of axial compression of proximal tibiae; and **(b)** comparisons between the predicted and ground truth reaction force values at various displacement levels.

attributed to the ViT-based model's use of Multi-head Self-Attention (MSA) (29), which enhances its ability to effectively aggregate slice-level features into volume-level representations.

### 3.4 Applicability of RoI-CT Imagery and MiceMediRec Datasets for Multi-modal Prognostic Assessment of Bone Mechanical Properties

Managing pathological fractures associated with bone metastases is crucial for preserving a patient's mobility and quality of life. In this context, we explore the applicability of our RoI-CT dataset to this critical scenario by assessing the mechanical competence of the bone at a future time point. We formulate this task as a multi-modal prognostic assessment of bone mechanical properties. First, the 3D CT scans in the RoI-CT imagery are used to train generative models to produce 3D CT scans of future frames, reflecting the progression of bone lesions. Next, the biological data in the MiceMediRec component is utilized to simulate the mechanical behaviors of axial compression of the proximal end of tibiae with metastatic osteolysis.

Three generative models are taken into account: 3D Generative Adversarial Networks (3D-GAN) (2), Temporal Variational Autoencoders (T-VAE) (33), and Spatial-Temporal Variational Autoencoders (ST-VAE), for generating future 3D CT scans. The quality of these generated scans is measured using PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity), and LPIPS (Learned Perceptual Image Patch Similarity) metrics. Note that a higher PSNR (and SSIM) value or a lower LPIPS value indicates better generation quality. Table 4 presents the quantitative results. We observed that all three methods achieve high-quality future CT generations. For instance, the 3D-GAN, T-VAE, and ST-VAE methods achieve SSIM values of 0.767, 0.817, and 0.860, respectively.

We take the qualitative results obtained using the ST-VAE model to visualize our results, as exhibited in Figure 6. In Figure 6a, the predicted 3D rendering volume is comparable to the ground truth, and the finite element (FE) simulations of axial compression on meshes constructed from the predicted bone structures closely resemble those constructed from the ground truth. Furthermore, Figure 6b illustrates the curves of the ground truth and predicted reaction force values at various displacement levels. It is evident that the predicted reaction force curve closely matches the ground truth, achieving an impressive $R^2$ value of 0.956. These results underscore the significance of our BoneMet data in the multi-modal prognostic assessment of bone mechanical properties with metastatic bone lesions.

### 3.5 Applicability of Sparse-Angle CT Reconstruction

Conventional CT reconstruction methods require 2D X-ray images from multiple angles. For example, using NRecon® software necessitates at least 260 X-ray images from different angles for accurate

Table 5: Quantitative evaluations of sparse-angle CT reconstruction

| Methods | PSNR (↑) | SSIM (↑) | FID (↓) | KID (↓) |
|---|---|---|---|---|
| PixelNeRF | 20.2 | 0.740 | 155.2 | 0.128 |
| MedNeRF | 30.2 | 0.810 | 91.5 | 0.092 |

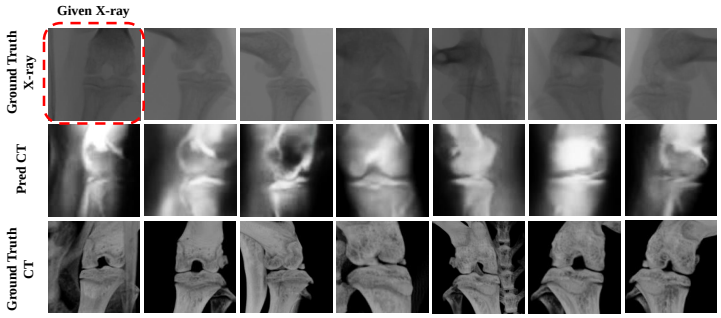



Figure 7: Illustration of sparse-angle 3D CT reconstruction from a 2D X-ray image by MedNeRF method. The first row also presents the given 2D X-ray image and X-ray images at other angles for reference. The second row shows the reconstructed 3D CT scans, while the third row displays the ground-truth 3D CT scans from the Recon-CT imagery.

reconstruction. This leads to prolonged exposure to ionizing radiation during the process of capturing multiple 2D X-ray images, incurring severe adverse effects on the patient's health. Here, we explore the applicability of our Rotation-X-Ray dataset to sparse-angle CT reconstruction, which can be validated by paired CTs in Recon-CT dataset. This task can significantly reduce radiation exposure by reconstructing 3D CT scans with continuous viewpoint rotations from a single 2D X-ray image.

We employed two NeRF-based methods: PixelNeRF (35) and MedNeRF (4), to reconstruct 3D CT scans. The reconstruction quality is measured by the PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), FID (Fréchet Inception Distance), and KID (Kernel Inception Distance) metrics. Notably, higher PSNR and SSIM values, along with lower FID and KID values, indicate better reconstruction quality. Table 5 presents the quantitative results of reconstructing 3D CT scans. It is observed that our dataset can support both NeRF-based methods for sparse-angle CT reconstruction. For example, the MedNeRF can achieve superb reconstruction quality, with a high PSNR (and SSIM) value of 30.2 (and 0.810) and a low FID (and KID) value of 91.5 (and 0.092). Moreover, we use the MedNeRF method to generate a complete set of CT projections within a full vertical rotation from a given single-view X-ray of a CT slice, with results shown in Figure 7. Despite the challenge of this task, the MedNeRF method produces high-quality reconstructions, demonstrating the applicability of our BoneMet dataset for sparse-angle CT reconstruction.

We also conducted additional experiments, including tibiae auto-segmentation, to validate the significance and broad applicability of our Recon-CT and Seg-CT dataset, and generalist biomedical AI diagnosis of metastatic breast tumor on bone, to demonstrate the applicability of our RoI-CT and MiceMediRec dataset. The details of these experiments are deferred to Sections B of the supplementary materials. Besides, the future potential applications of the BoneMet dataset are discussed in Sections D of the supplementary materials.

## 4 The BoneMet Package

In addition to our BoneMet dataset, we also develop the BoneMet package, including three types of APIs: 1) CT Image Segmentation, 2) CT Image Registration, and 3) RoI-based CT Image Cropping), at the Python Package Index(PyPI), for public release to facilitate our dataset's ease access. The details of three APIs and their usage examples are deferred to Section C of supplementary materials.

## 5 Conclusions

This paper introduces the BoneMet dataset, the first large-scale meticulously curated collection of well-organized CT and X-ray images in angles, positions and time points designed specifically for the

diagnosis and prognosis of BTBM disease. It includes high-resolution images that are well-organized across spatial and temporal dimensions, allowing for an in-depth analysis of bone responses to tumor progression and treatment. The integration of detailed imaging with medical metadata enhances the utility of the BoneMet dataset, enabling the development and validation of advanced deep learning models. This dataset is also invaluable for developing high-quality CT reconstruction techniques from sparse X-ray data and for performing 3D segmentation. Our extensive experimental evaluations confirm that the dataset is compatible with various deep learning approaches. Alongside the dataset, we have developed the BoneMet Package to assist researchers and practitioners in automatic breast tumor bone metastasis medical images processing such as 3D CT images segmentation, registration and region of interest (RoI) selection, as well as constructing their own deep learning models. While the primary aim of the BoneMet dataset is to advance deep learning models in disease diagnosis, management and medical image processing techniques, its potential applications extend across medical imaging, oncology, and computational pathology. We believe that our BoneMet dataset will be a significant valuable asset to the fields of deep learning, medical radiology, orthopedics and oncology, spurring further interdisciplinary research at the intersection field of healthcare and artificial intelligence.

## 6 Broader Impacts

The use of preclinical models is well justified and necessary in cancer research. As elaborated in the "Guidelines for the welfare and use of animals in cancer research" (32), animal experiments remain essential to understand the fundamental mechanisms driving malignancy and to develop individualized molecularly targeted cancer therapies for humans. With the explosion of newly created genetic animal models, the various genes, signaling pathways, and risk factors that give rise to cancer dependencies and vulnerabilities have been identified, and the responses to cancer treatments can be comprehensively tested. Thus animal models and testing not only extend our genetic, molecular, and holistic understanding of cancer but also gather necessary safety and efficacy information that is required to introduce new drugs and diagnostic tools into clinics. Specifically, our dataset was collected from mice implanted with triple-negative breast cancer, which is the most aggressive among all subtypes of breast cancer with limited treatment options. This dataset takes advantage of our animal models, which mimics the full course of the disease development from microscopic metastasis lesions to macroscopic metastasis (whole bone failure). Currently, we have collected 2D and 3D images of the whole bone and tumor-affected regions at five-time points, along with the biometrics of the animal subjects such as their age and body weight. The Dataset is designed to expand and include additional biochemical data such as tumor burden, sera biomarkers, gene transcripts, and histological results. Preclinical research in the cancer field as well as our own studies support the similarities between the mice and humans in terms of the fundamental mechanisms driving cancer spread and growth in the bone environment. Our specific goal is to reveal specific imaging biomarkers associated with tumor growth and bone metastasis, and further make it possible to translate into clinical applications, enabling better understanding, detection, and treatment of bone metastases in human breast cancer.

The BoneMet dataset engaged with investigators, clinicians, and policymakers regarding the clinical and research needs regarding cancer bone metastasis diagnosis and treatment. By publishing our BoneMet dataset, we envision that more stakeholders can benefit from full access to the dataset, which could be used to develop early diagnostic tools for breast cancer metastasis with improved accuracy and sensitivity. Clinically, the disease progression is much faster in mice than in humans, which allows the capture of images representing the full course of the disease progression. Such a dataset contains a large span of the disease severity and granular longitudinal changes, which could be used to model human cross-sectional data covering different disease stages. The dataset and the insights it generates could play a crucial role in advancing AI applications, such as large vision models and multimodal approaches with temporal and spatial alignment, for the early diagnosis and prognosis of breast tumor bone metastasis, particularly in enhancing the predictive accuracy of models used in clinical settings when the disease progression is relatively slower. As we continue to refine and expand our dataset, we aim to provide a robust resource for better understanding, detection, and treatment of bone metastases in human breast cancer, and further support future innovations in deep learning studies and clinical practices.

## REFERENCES


[1] Medpix medical image database. http://rad.usuhs.mil/medpix/medpix.html, 2024. Free Online Medical Image Database.

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.

[3] Gary JR Cook, Gurdip K Azad, and Vicky Goh. Imaging bone metastases in breast cancer: staging and response assessment. *Journal of Nuclear Medicine*, 57(Supplement 1):27S–33S, 2016.

[4] 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.

[5] Colleen M Costelloe, Eric M Rohren, John E Madewell, Tsuyoshi Hamaoka, Richard L Theriault, Tse-Kuan Yu, Valerae O Lewis, Jingfei Ma, R Jason Stafford, Ana M Tari, et al. Imaging bone metastases in breast cancer: techniques and recommendations for diagnosis. *The lancet oncology*, 10(6):606–614, 2009.

[6] Steven S Coughlin and Donatus U Ekwueme. Breast cancer as a global health concern. *Cancer epidemiology*, 33(5):315–318, 2009.

[7] Yusuf E Erdi, John L Humm, Massimo Imbriaco, Henry Yeung, and Steven M Larson. Quantitative bone metastases analysis based on image segmentation. *Journal of Nuclear Medicine*, 38(9):1401–1406, 1997.

[8] Anna Fantozzi and Gerhard Christofori. Mouse models of breast cancer metastasis. *Breast Cancer Research*, 8:1–11, 2006.

[9] CSB Galasko. Diagnosis of skeletal metastases and assessment of response to treatment. *Clinical Orthopaedics and Related Research®*, 312:64–75, 1995.

[10] Margery Gass and Bess Dawson-Hughes. Preventing osteoporosis-related fractures: an overview. *The American journal of medicine*, 119(4):S3–S11, 2006.

[11] Theresa A Guise. The vicious cycle of bone metastases. *Journal of Musculoskeletal and Neuronal Interactions*, 2(6):570–572, 2002.

[12] Tsuyoshi Hamaoka, John E Madewell, Donald A Podoloff, Gabriel N Hortobagyi, and Naoto T Ueno. Bone imaging in metastatic breast cancer. *Journal of Clinical Oncology*, 22(14):2942–2953, 2004.

[13] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis Langlotz, Serena Yeung, Nigam Shah, and Jason Alan Fries. Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

[14] Nancy L Keating, Mary Beth Landrum, Edward Guadagnoli, Eric P Winer, and John Z Ayanian. Surveillance testing among survivors of early-stage breast cancer. *Journal of Clinical Oncology*, 25(9):1074–1081, 2007.

[15] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*, volume 12350, pages 491–507, 2020.

[16] Ajith AK Kumar. Mortality prediction in the icu: The daunting task of predicting the unpredictable. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 26(1):13, 2022.

[17] Chaofan Li, Mengjie Liu, Jia Li, Weiwei Wang, Cong Feng, Yifan Cai, Fei Wu, Xixi Zhao, Chong Du, Yinbin Zhang, et al. Machine learning predicts the prognosis of breast cancer patients with initial bone metastases. *Frontiers in Public Health*, 10:1003976, 2022.

[18] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[19] Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, et al. Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5784, 2023.

[20] Jiaming Liu, Rushil Anirudh, Jayaraman J Thiagarajan, Stewart He, K Aditya Mohan, Ulugbek S Kamilov, and Hyojin Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10508, 2023.

[21] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.

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

[23] Nikolaos Papandrianos, Elpiniki Papageorgiou, Athanasios Anagnostis, and Anna Feleki. A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. *Applied Sciences*, 10(3):997, 2020.

[24] DavidJ Perez, Jonathan Milan, HT Ford, JamesS Macdonald, TrevorJ Powles, JC Gazet, V Ralph Mccready, and R Charles Coombes. Detection of breast carcinoma metastases in bone: relative merits of x-rays and skeletal scintigraphy. *The Lancet*, 322(8350):613–616, 1983.

[25] Jian Rong, Siyang Wang, Qiue Ding, Miao Yun, Zhousan Zheng, and Sheng Ye. Comparison of 18fdg pet-ct and bone scintigraphy for detection of bone metastases in breast cancer patients. a meta-analysis. *Surgical oncology*, 22(2):86–91, 2013.

[26] Roman Shyyan, Shahla Masood, Rajendra A Badwe, Kathleen M Errico, Laura Liberman, Vahit Ozmen, Helge Stalsberg, Hernan Vargas, László Vass, Global Summit Diagnosis, and Pathology Panel. Breast cancer in limited-resource countries: diagnosis and pathology. *The breast journal*, 12:S27–S37, 2006.

[27] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning (ICML)*, pages 10096–10106, 2021.

[28] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeruIPS)*, pages 5998–6008, 2017.

[30] Shubo Wang, Shaopeng Pei, Murtaza Wasi, Ashutosh Parajuli, Albert Yee, Lidan You, and Liyun Wang. Moderate tibial loading and treadmill running, but not overloading, protect adult murine bone from destruction by metastasized breast cancer. *Bone*, 153:116100, 2021.

[31] Murtaza Wasi, Tiankuo Chu, Rosa M Guerra, Rory Kooker, Kenneth Maldonado, Xuehua Li, Chun-Yu Lin, Xin Song, Jinhu Xiong, Lidan You, et al. Mitigating aging and doxorubicin induced bone loss in mature mice via mechanobiology based treatments. *Bone*, 188:117235, 2024.

[32] P Workman, EO Aboagye, F Balkwill, A Balmain, Gail Bruder, DJ Chaplin, JA Double, J Everitt, DAH Farningham, MJ Glennie, et al. Guidelines for the welfare and use of animals in cancer research. *British journal of cancer*, 102(11):1555–1577, 2010.

[33] Wei Xiong, Neil Yeung, Shubo Wang, Haofu Liao, Liyun Wang, and Jiebo Luo. Breast cancer induced bone osteolysis prediction using temporal variational autoencoders. *BME frontiers*, 2022.

[34] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

[35] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[36] Yiqiu Zhang, Beilei Li, Haojun Yu, Junyi Song, Yuhong Zhou, and Hongcheng Shi. The value of skeletal standardized uptake values obtained by quantitative single-photon emission computed tomography-computed tomography in differential diagnosis of bone metastases. *Nuclear Medicine Communications*, 42(1):63–67, 2021.

[37] Zhen Zhao, Yong Pi, Lisha Jiang, Yongzhao Xiang, Jianan Wei, Pei Yang, Wenjie Zhang, Xiao Zhong, Ke Zhou, Yuhao Li, et al. Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. *Scientific reports*, 10(1):17046, 2020.

[38] Zhuo Zhi, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. Multimodal diagnosis for pulmonary embolism from ehr data and ct images. In *Engineering in Medicine & Biology Society (EMBC)*, pages 2053–2057, 2022.