Synthetic Data Generation for Small Data Orthopedic Radiology Using Latent Diffusion Models

Max Van Veen
VVNMAX002@myuct.ac.za
University of Cape Town
Cape Town, Western Cape, South Africa

Abstract

Manual orthopedic X-ray diagnosis suffers from a combination of time cost and human error. However, the lack of large and high-quality data sets due to privacy concerns has led to existing medical data sets being small and containing adversarial or messy data. This has hindered the development of suitable pathology diagnosis tools. Several architectures have been proposed for the purposes of image generation, with the most prominent being Generative Adversarial Network-based and diffusion-based models. However, few of these have been applied to medical imaging problems and large diffusion-based models such as Flux.1-dev and Stable Diffusion XL have been avoided due to high fine-tuning costs. Thus, the purpose of this study was to critically analyze three older Generative Adversarial Network approaches and two state-of-the-art latent diffusion approaches to image generation with regard to their efficacy in learning the underlying patterns in elbow radiograph images to generate suitably realistic images for training classification models. This study presents a mask-based fine-tuning method for SDXL and FLUX that has been validated in terms of the Structural Similarity Index Measure, Multiscale Structural Similarity Index Measure, and Maximum Mean Discrepancy metrics and represents a state-of-the-art approach for the creation of synthetic medical data. The primary findings demonstrated that fine-tuned latent diffusion-based approaches, specifically the Flux.1-dev model, were the most suitable for the task of radiograph image generation due to several improvements over Generative Adversarial Network approaches. This has resulted in a technical contribution to the field of synthetic medical data generation with a cutting-edge approach that scores well across the aforementioned metrics.

1 Introduction

With the rise of statistical-based machine learning algorithms and the prominence of the neural network as a universal function approximator, these algorithms have seen a large proliferation in terms of the number of subarchitectures and widespread use in a wide variety of fields. One such field is that of Computer-Aided Diagnosis (CAD), where machine learning algorithms such as the Convolutional Neural Network (CNN) have been used to identify pathologies present in a variety of medical imaging fields [11][33][1][50]. Due to the ability of machine learning algorithms to find and exploit underlying patterns in medical data, relative success has been seen in accurately diagnosing various pathologies, with several models achieving accuracy scores in the high 80% to low 90% range [11][50]. Despite this success, these algorithms have not reached the accuracy scores of 95%-97% for human radiologists [58]. However, these approaches offer methods to reduce the number of human errors, which account for approximately 40 million diagnostic errors annually per year [24]. These misdiagnoses result in unnecessary deaths, unintended injuries or complications, wasteful medical expenditure, and malpractice lawsuits [24]. Therefore, these machine learning algorithms have the potential to save lives, money, and time.

Several issues arise from the need to train CAD models, the most prominent of which is that these algorithms require large, clean, and evenly distributed data sets in order to make optimal use of the information and produce accurate diagnoses. Due to policy and regulatory frameworks such as the US Health Insurance Portability and Accountability Act (HIPAA) [46] and the General Data Protection Regulation (GDPR) [5], which were created to control the dissemination of medical data and ensure the privacy of confidential patient information, these regulations have led to issues such as medical data sets that are often small, unevenly distributed, and contain adversarial data. This is due to the high costs and difficulty associated with the anonymization process of medical data and the laborious procedures involved with the sharing of medical data with third countries [15].

Therefore, several image generation machine learning algorithms based on the Generative Adversarial Network (GAN), such as the Wasserstein GAN (WGAN) [3] and the Auxiliary Classifier GAN (AC-GAN) [53] have been used for the purposes of anonymization of medical data and extension of data sets [17][13][19]. With a notable exception being the Twin Auxiliary Classifiers GAN (TAC-GAN), which is theorized to be an improvement over the AC-GAN architecture [35]. A separate branch of image generation machine learning algorithms is the latent diffusion-based models [27], namely the Stable Diffusion family of models such as the Stable Diffusion XL 1.0 (SDXL) [16] and the Flux.1-dev (FLUX) models [51]. The latent diffusion approach to image generation is considered a state-of-the-art technique and has been shown to outperform both GAN-based and Variational Autoencoder (VAE)-based models for image generation tasks [27] [9] [6]. Diffusion-based models have seen success in their use for the generation of synthetic Computer Tomography (CT) scans, Magnetic Resonance Imaging (MRI) scans, and microscopic images [1] [33]. Despite the images showing promising quality, incorporating them into the training data either worsened the performance of the classification models or did not enable the classification model to achieve human-level performance. Furthermore, minimal effort has been made to fine-tune existing models such as SDXL and FLUX due to the large costs associated with training these models [41] [51]. A notable feature of these models is mask-based generation, which offers both context of the image to the models and a reduction in the amount of image pixels required

to learn. This, in combination with text prompts provided to the models, simplifies the image-generating process for the models and allows them to be fine-tuned for the generation of specific areas in medical images where specific pathologies may occur. This has the potential to allow for more accurate and focused synthetic images.

This study begins with key examples of previous GAN-based approaches to the generation of synthetic medical images, newer diffusion-based approaches to the generation of synthetic medical images, state-of-the-art latent diffusion-based approaches to the generation of images, and descriptions of the three key metrics chosen for model evaluation, in Section 2. Thereafter, the general experimental methodology is described in Section 3. Finally, the results are presented in Section 4 and discussed in Section 5.

1.1 Research Question

Does FLUX outperform SDXL and do these latent diffusion-based methods outperform WGAN-GP, AC-GAN, and TAC-GAN in terms of their efficacy in learning the underlying patterns in elbow radiograph images to generate suitably realistic images for training classification models?

This question addresses several key issues in current state-of-the-art research, namely, the lack of existing studies which fine-tune the SDXL and FLUX models, the lack of application of the SDXL, FLUX, and TAC-GAN models into synthetic medical data generation, the low data problems experienced during the training of CAD models [11][58] due to restrictions on the compilation and sharing of large medical data sets [46][5][15], and the issue of current synthetic medical data not being of the same quality as real data [1][17][19][33]. Therefore, this is an important and relevant question in the context of synthetic medical data generation and, in answering it, has the potential to aid the development of human performance level CAD tools for the purposes of reducing human error in diagnoses, potentially saving money, time, and human lives. The efficacy of each model will be measured by measuring SSIM [45], MS-SSIM [59], and MMD [43] on the output synthetic radiographs of each model and plotting the curves for each of these metrics as the fine-tuning or training epoch increases. It should be noted that this study does not attempt to justify the superiority of latent diffusion-based models over GAN-based models, as this has been extensively covered in Section 2; rather this paper aims to determine whether fine-tuning large, pre-trained latent-diffusion models may allow for more stable learning and better applicability to medical radiographs than GAN-based methods. Please, see Table 2 for the experiments chosen to analyze the research question.

2 Related Works

This section covers both the older GAN-based and state-of-the-art latent diffusion-based image generation techniques. Architectural improvements are discussed and the superiority of latent diffusion in image generation over GAN-based methods is justified. Two key latent diffusion models are identified and performance comparisons between the two models are discussed. Thereafter, the chosen metrics for image evaluation are described.

2.1 GAN-Based Image Generation

Prezja et al. [17] proposed a WGAN with Gradient Penalty (WGAN-GP) model for the generation of knee osteoarthritis X-ray images which produced images of sufficient quality such that a panel of experts could distinguish between real and synthetic images with an accuracy of only 61.35%. However, images were generated at only 210x210 pixels, which may have influenced the ability of the experts to accurately classify images as real or synthetic, and when synthetic images were used to train a classification model, the classification accuracy dropped by 3.79% compared to purely real images. Indicating that the generated images were not of the same quality as the real images.

Sun et al. [19] proposed an AC-GAN model for the generation of synthetic MRIs for vertebral units, which were evaluated in terms of diversity and fidelity. The synthetic images produced by the model showed good generalizability and little overfitting [19], indicating good diversity. However, the classification model trained on synthetic images performed worse than when trained on purely real images, indicating poor realism and fidelity.

TAC-GAN [35] improved the base AC-GAN architecture by adding a second auxiliary classification model (the twin classifier). This addressed a core logical error present within AC-GAN, which ignored the negative conditional entropy when training. This improvement allowed TAC-GAN to theoretically overcome the low intra-class diversity problem present in AC-GAN and reduced the tendency of the model to mode collapse [35]. At the time of writing, there are no known implementations of TAC-GAN within a medical context.

2.2 Diffusion-Based Image Generation

Marioriyad et al. [9] proposed that diffusion-based models outperform VAE-based models in compositional generation ability by evaluating several state-of-the-art diffusion-based models and two state-of-the-art autoregressive-based models on the T2I-CompBench data set [30]. To test compositional generation ability, a variety of metrics, such as CLIP similarity [23], BLIP-VQA (to measure attribute binding capabilities) [25], and UniDet (to measure relational positions and count the number of objects in synthetic images) [60], were used. It was found that diffusion models generally outperformed VAEs in correctly aligning visual images with the textual prompts provided, especially when the textual prompts were complex [9].

Dhariwal & Nichol [6] proposed that diffusion-based models outperform GAN-based models in generational sample quality and stability by using a variety of metrics such as Frechet Inception Distance (FID), Spatial FID (sFID), precision, and recall. Diffusion-based models generally outperformed GAN-based models in these metrics, with consistently lower FID and sFID scores, implying that diffusion-based models produced images that were more similar to real images in terms of quality and diversity, and had consistently higher precision and recall scores, indicating better fidelity and diversity (better data coverage) of generated samples [6].

Nguyen et al. [33] proposed a lightweight diffusion-based model for Computer Tomography (CT) chest scans of the SARS-CoV-2 pathology, which showed promising generation quality when qualitatively analyzed. In addition, al Nomaan Nafi et al. [1] proposed a separate set of diffusion-based models for brain tumor MRI scans, Acute Lymphoblastic Leukemia (ALL) microscopic images, and chest CT scans of the SARS-CoV-2 pathology, respectively. These models were used to generate synthetic data sets of 1700 brain tumor MRI, 1000 microscopic ALL, and 1500 Sars-CoV-2 CT scans. Eight state-of-the-art classification models were trained on the resulting synthetic data set, with ResNet-50 [29] achieving the highest accuracy of 78.24% for Sars-Cov-2 CT scans, VGG-19 [57] achieving an accuracy of 86.46% for brain tumor scans, and DenseNet-121 [18] achieving an accuracy of 91.38% for the ALL images. However, the classification models were not trained on purely real images, and the accuracy scores were not as high as those of human specialists [58].

2.3 Stable Diffusion

Rombach et al. [41] identified that diffusion-based models surpassed other methods such as GAN-based models in terms of quality and stability. However, a key issue was the significant increase in computational resources required to train and conduct inference on a diffusion-based model. Therefore, the technique known as latent diffusion was proposed as a solution to the significant computational costs while preserving important semantic and perceptual details. The Stable Diffusion family of models leveraged this approach to see a significant reduction in computational cost while maintaining competitive performance in unconditional image generation, text-to-image synthesis, resolution up-scaling, and image inpainting in terms of FID, IS, precision, and recall scores [41]. At the time of writing, there are no known implementations of fine-tuning a Stable Diffusion model for the purpose of generating synthetic medical data.

2.4 FLUX

The FLUX family of models is a set of new text-to-image latent diffusion models developed by Black Forest Labs [51]. The models mark an improvement over other latent diffusion-based methods, such as the Stable Diffusion family. At the time of writing, no technical report existed for FLUX. However, Marioriyad et al. [9] propose that FLUX employs a hybrid architecture consisting of numerous state-of-the-art diffusion-based architecture techniques, such as multimodal [39] and parallel [34] diffusion transformer [54] blocks operating within a flow matching [44] framework, rotary positional embeddings [26], and parallel attention layers [34] to achieve state-of-the-art image generation capabilities. At the time of writing, the FLUX family of models had not been applied to synthetic medical data generation because of the expensive fine-tuning costs and the recency of the model.

The FLUX-dev variant of the FLUX family was shown to be superior to the Stable Diffusion family by Marioriyad et al. [9]. In addition to testing the latent diffusion-based models, DALL-E3 [22], a leading closed source transformer-based model, was tested, and FLUX demonstrated similar performance while outperforming Stable Diffusion on every metric [9]. In addition to this, a 1.58 bit quantization of FLUX [14] showed better performance than Stable Diffusion XL despite less accurate weights on both the T2I CompBench [30] and GenEval [47] data sets.

2.5 Image Evaluation Metrics

SSIM [45] measures the difference between images by computing the luminance, contrast, and structure of each image. The luminance, contrast, and structure capture the differences in image brightness, pixel intensity, and pixel spatial interdependencies, respectively. This allows the SSIM metric to offer a more human-visible evaluation of image quality, in contrast to traditional pixel-based similarity metrics [56], while maintaining good performance due to simple mathematical formulation and ease of parallelization. This has made the metric a popular and widely used metric in the field of computer vision [20], making it an important metric to report. However, SSIM is sensitive to spatial scale selection [28][10]. A widely used variant of SSIM that addresses this sensitivity is MS-SSIM [59], which extends the SSIM metric by creating weighted mean SSIM scores across several resolutions. This allows many different scales to be captured in the final score, a key factor present in the data set which had a wide variety of resolutions present.

MMD is a distribution distance metric that measures the proximity of two probability distributions by mapping the distributions to reproductive kernel Hilbert spaces, where comparisons are more flexible due to the kernel being freely chosen, and trying to minimize the mean deviance between the source and target domains [43][8]. Thus, reducing the difference between the probability distributions. A key advantage is that MMD makes minimal assumptions about the input data, which enables the metric to be used in a wide variety of applications [4]. This flexibility is useful for adversarial medical data.

2.6 Summary

In summary, several issues were identified with previous approaches to the creation of synthetic medical data. Most important of which is the fact that previous approaches see a decrease in classification accuracy when training CAD models on pure synthetic data compared to pure real data and that previous approaches were unable to increase CAD classification accuracy to human-level accuracy even when data mixing was employed. Therefore, a clear need for models that can produce more realistic and diverse synthetic images arises. One potential solution to this is large pre-trained latent-diffusion models such as the SDXL and FLUX models, which were shown to be experimentally and theoretically superior in image generation by previous research, when compared to VAE- and GAN-based approaches.

3 Methods

This section discusses the methods performed to address the aforementioned issues with previous approaches to synthetic medical data generation by describing the data preprocessing and splitting procedure, the image evaluation metrics chosen, the GAN-based

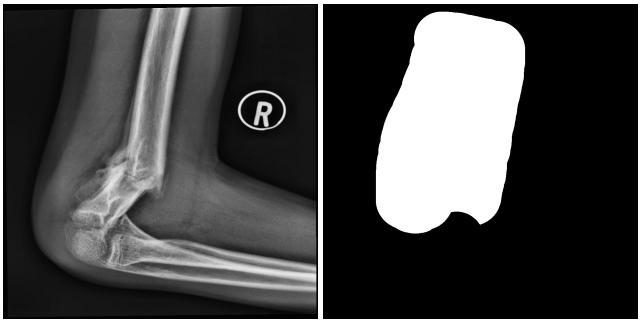

Figure 1: Figure showing a LAT radiograph image for an elbow exhibiting both supracondylar fracture and soft tissue swelling (left) and the image mask drawn for the supracondylar fracture, drawn around the humerus (right). The generated label for this image read as "A lateral x-ray of an elbow displaying soft tissue swelling, supracondylar fracture". Note that the mask was specifically drawn to cover the supracondylar fracture and a second separate image mask had been drawn over the entire elbow for the soft tissue swelling pathology.

method training procedures in general, and justifying the hyperparameter choices for the SDXL and FLUX models. Please see the appendix for the GAN-based model architectures and their training hyperparameters (Table 4, Table 5, and Table 6).

3.1 Data Set Preprocessing

In computer vision, applying transformations to images before they are used to train a model is a common and crucial step in the learning process. Contrast Limited Adaptive Histogram Equalization (CLAHE) is a commonly applied image transformation that is often applied to both machine- and human-based medical imaging tasks due to the effective local image contrast enhancement of the transformation [36] [12] [32] [17]. CLAHE is able to reveal fine details that have been obscured by poor illumination or low contrast inherent to radiograph images while avoiding noise amplification. The method has also shown good results in the field of medical classification, particularly knee osteoarthritis radiographs [17], brain MRIs [48], and diabetic retinopathy images [36]. Therefore, CLAHE was chosen as a suitable image transformation and was applied to the entire data set of 2783 elbow radiographs. In addition to CLAHE, an automated cropping check was performed to resize any images that incorporated additional unnecessary padding around the radiographs, images were checked for color inversion, all images were resized to 1024x1024 for uniformity, and anteroposterior (AP) and lateral (LAT) radiographs were separated to ensure data set uniformity and allow smoother training of the models.

After the above steps were performed on the elbow radiographs, the image labels were automatically generated in natural language to retrain the latent diffusion model text encoders, and the image masks were hand-drawn according to the general area in which each pathology occurred for a total of 3017 unique masks (see Figure 1 for an example). The image labels followed the format of "A [LAT/AP] x-ray of an elbow displaying [pathology 1], [pathology 2], ...". This meant that for a single elbow radiograph with several pathologies present, a distinct mask was drawn for each pathology, regardless of any pathologies that share the same area. For example, radiographs that showed only joint effusions and olecranon fractures would have two distinct masks that encompass the elbow joint. This allowed for more efficient use of data, due to the natural variations present in the masks for the same area of the same radiograph and allowed the models more training runs over the more complex multi-pathology masks. A crucial consideration that was made due to the low data situation. The masks varied in terms of size, shape, the amount of soft tissue captured, and the amount of background, non-tissue captured. Non-tissue portions of the images were chosen to be included as a subset of the image masks because they forced the models to learn and generate the edges of the elbow, rather than solely the interior tissue. The following is a list of pathologies organized by where the masks were drawn for radiographs with that pathology.

3.1.1 Joint Area.

- Joint effusion
- Medial epicondyle displaced
- Lateral epicondyle displaced
- Olecranon fracture
- Radial head fracture
- Radial head subluxation

3.1.2 Humerus (Upper Arm).

- Distal humerus fracture
- Supracondylar fracture

3.1.3 Radius and Ulna (Forearm).

- Proximal ulnar metaphysis fracture
- Proximal radial fracture

3.1.4 Whole Arm.

- Elbow dislocation anterior
- Elbow dislocation posterior
- Soft tissue swelling

3.1.5 Normal Elbows. Normal elbows were any elbows without pathologies present; in other words, radiographs displaying a healthy elbow. Masks were drawn in an even distribution of the joint, upper arm, forearm, and whole arm for normal elbows.

Finally, the data set was divided into a training and test data set, with 30% of the total data being used for an out-of-sample test data set and the remainder being used for the training data set. Crucially, the test images were sampled such that multi-pathology images were not repeated in the training data set by having any multi-pathology image in the test data set count for multiple pathology classes and completely removing it from the test data set. Please, see Table 3 for a breakdown of the train and test data sets with their relative pathology counts. Of particular interest are the pathologies that suffer from the lowest data, namely elbow dislocation posterior, proximal ulnar metaphysis fracture, radial head fracture, and radial head subluxation. These low-data pathologies are due to the

rareness of these types of injuries in the real world and are particularly difficult for both classification models to accurately classify and generative models to accurately replicate due to the small number of them present in the training data set. These pathologies were treated as equal in the metric calculation, and a full analysis of individual pathology scores was not feasible due to the page limit.

3.2 Metrics

The metrics chosen for the quality evaluation were SSIM, MS-SSIM, and MMD. These metrics were chosen because (1) SSIM is a universal, widely applicable metric that has been widely reported in previous literature [20]. (2) MS-SSIM addresses the scaling sensitivity issue present in SSIM [59], a key feature for an accurate comparison between the GAN-based and latent diffusion-based models that generated images at different resolutions. And (3) MMD is better equipped than the more commonly used Fréchet Inception Distance (FID) for small medical data sets with non-normally distributed image features, primarily due to the fact that MMD does not make assumptions about the shape of the data distributions as opposed to FID [4]. Alternative metrics such as Feature Similarity Index Measure were observed to never score synthetic-real image pairs below 0.5 (even when comparing random noise to real images), meaning that this metric is unsuitable for showing whether a generative model that starts with random noise is learning (such as a latent diffusion model), and Information Content Weighted Structural Similarity Index and Peak Signal-to-Noise Ratio both saw large score changes for minor image modifications. MMD was computed using a radial basis function kernel and using a pretrained ResNet-50. In addition to these metrics, 50 random images were subjected to manual review from a non-expert for each of the models (50 per 20 epochs for the diffusion models) to identify obvious image generation issues such as blurriness. A more comprehensive and objective manual review was not possible due to resource limitations and ethical concerns. The selected metrics were computed between the real images in the test data set and the synthetic image pairs generated based off the real images in the test data set. Only one synthetic image per real testing sample was generated in order to ensure uniformity. A key note and potential limitation of this study was that no extensive statistical tests were not undertaken; however, this was infeasible due to time constraints and the large amount of time required to implement the GAN-based models and fine-tune and conduct inference on the latent diffusion-based models. A classification model was not trained on synthetic data because distribution-based metrics require a large test data set for objective evaluation and for compatibility with the GAN-based methods. This reduction in training data would lead to poorer generative performance. Therefore, it was viewed that future research should train a classification model, where the test data set can be formed in a manner that makes better use of the available data by complementing the mask generation ability of the latent diffusion-based models. An example of such an approach is to generate synthetic fractures off the normal elbows in order to test the models.

3.3 GAN-Based Methods

Following the implementation of Prezja et al. [17], a WGAN-GP was developed due to the promising performance of the model for knee osteoarthritis radiographs. The model was adapted to be trained on elbow radiographs in a non-masked, non-labeled fashion. This is due to the architecture of the WGAN-GP, which does not inherently support masked or labeled data. This meant that a unique WGAN-GP had to be trained and sampled for each pathology. In addition to this, separate models were trained for LAT and AP images, with right-oriented radiographs being mirrored across the Y axis for additional data consistency. Since WGAN-GP represented a substantial improvement in the image-generating capabilities of the vanilla GAN [3][21], it was decided that this model would serve as a suitable baseline model from which the other models would be compared for hyperparameter choices. The WGAN-GP was a suitable choice because it is widely researched, is easier to train than a vanilla GAN due to the reduced chance of mode collapse and greater training stability, and the other models were expected to have outperformed the simpler WGAN-GP model in terms of image generation capabilities. Please, see Table 4 for a breakdown of the model architecture and hyperparameters.

Due to issues experienced with mode collapse around epoch 30 when replicating the AC-GAN architecture implemented by Sun et al. [19], the AC-GAN was developed following a mixed implementation of Sun et al. [19] and Dhawan and Nijhawan [7]. The TAC-GAN was derived from the AC-GAN model to ensure a fair comparison, and following the implementation by Gong et al. [35], the twin auxiliary classifier was added. These models were trained in two stages: (1) a hyperparameter search stage where the hyperparameters were tuned to achieve the most realistic images according to the non-expert evaluation, and (2) hyperparameters were selected based on the previous results and the final models were trained using image labels but not masks. This process was necessary because both AC-GAN and TAC-GAN are sensitive to input and output resolution and class distributions in data sets with high intraclass overlap [53] [35], such as medical data sets with multi-pathology images.

All GAN-based models were developed using the Pytorch framework, trained for 4000 epochs on Nvidia L40 graphical processing units, and checkpointed and sampled on the test set every 500 epochs. TAC-GAN required 52 hours, AC-GAN required 48 hours, and each WGAN-GP model required 32 hours of training. The images were trained and generated at 256x256 pixels because larger values led to a significantly higher chance of mode collapse among the GAN-based models. Subsequently, the synthetic images were resized to 1024x1024 pixels using bilinear interpolation and passed on to the SSIM, MS-SSIM, and MMD metrics for quality evaluation. The generated images were resized to avoid information loss in the test images, to standardize metric input size across generation methods, and to ensure a more even comparison with the diffusion-based models. However, it should be noted that this may have artificially inflated the GAN-based SSIM and MS-SSIM metrics because bilinear interpolation smooths noise and increases local pixel correlations.

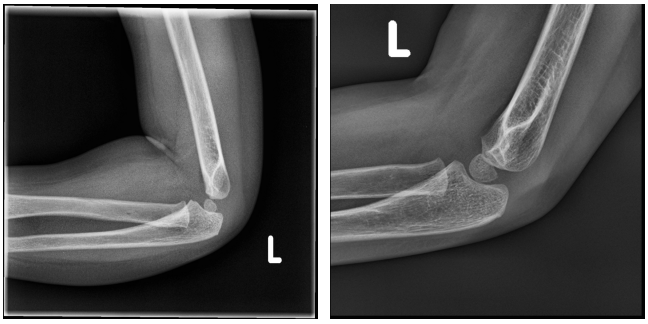

Figure 2: A figure showing two sample normal (healthy) LAT elbows. Despite the fact that both of these elbows are left-oriented, the right image is slightly brighter than the left image, the elbows have different amounts of soft tissue, different bend angles, different bone orientations, and joint locations of the elbows are different.

This should have been offset by the fact that bilinear interpolation was a conservative upsampling method that preserves relative structural content without adding high-frequency hallucinations that would have biased the metrics.

3.4 Diffusion-Based Methods

Both the SDXL and FLUX models were fine-tuned using the open source OneTrainer tool [52]. Respectively, the inpainting and fill (masked) versions of the models were chosen because fine-tuning the models based on unconditional image generation would have resulted in an inability to extract synthetic labels, which is crucial for evaluating the ability of a model to generate specific pathologies, and pure text-to-image (prompted) generation would not have benefited from individual image context. This would have caused the model to potentially produce less realistic images due to noise present in the data, such as poorly positioned body parts or unevenly distributed body part positions. An example of such was right- and left-oriented elbows, where WGAN-GP required an additional check to flip right-oriented radiographs (training dedicated right- and left-oriented models was infeasible due to the small size of the data set). However, this was only able to address the extreme case of completely inverted radiographs, and not more minor variations such as bone orientation, joint position, elbow bend angle, amount of soft tissue, or image brightness. Please see Figure 2 for an example of this, where it is shown that the two elbows have different overall brightnesses, amounts of soft tissue, bend angles, bone orientations, and joint positions. Furthermore, numerous radiographs had unrelated body parts present, such as another arm or the shoulder and ribs of the patient. By including the overall image context, as was with mask-based image generation, the model is shown crucial information which would allow it to determine the overall image brightness, certain bone orientations, and the amount of soft tissue expected for the synthetic image. Therefore, generating a higher quality image. An additional benefit of mask-based fine-tuning is that the model is allowed to focus on the specific areas that are most important for generating realistic synthetic pathologies, such as the joint area for the joint effusion pathology. In contrast to unmasked fine-tuning, as is the case for all of the GAN-based models, where these details would have to be learned alongside unimportant details such as the positions of the lead markers (the 'L' and 'R' letters used to indicate the left or right side of a patient's body, which the GAN-based models visibly struggled with). Thus, mask-based models were selected for fine-tuning and testing. However, it should be noted that no tests were conducted for the comparison of the mask-based and pure text-to-image models due to time constraints and the length of time required to fine-tune both models.

At the time of writing, the author was unaware of any published studies discussing and testing different fine-tuning hyperparameter settings for either the SDXL or FLUX models. Therefore, the justifications for the hyperparameters chosen in Table 7 were entirely based on the consensus of the community on the choices of hyperparameters. Experiments were performed for critical hyperparameters such as the learning rates, data types, batch sizes, and optimizers, but due to time constraints and the length of time required to fine-tune the models, extensive tests were infeasible. A large number of hyperparameters were kept the same between SDXL and FLUX to ensure a fair comparison and to reduce the amount of fine-tuning undergone. The learning rates were tested at the values of 1×10^{-3} and 5×10^{-4}; however, both showed good metric scores at the level of 20 epochs but began to drop thereafter. This was an indication that the learning rate was too high and was causing training instability. A learning rate of 1×10^{-5} showed good metric curves and was a good value according to community consensus. Smaller learning rates were not explored due to time constraints and the need for fast convergence due to the small training data set. The data types did not affect the model scores in any meaningful way; however, since SDXL was a smaller model, it was able to be trained at float32 precision without a significant increase in training time. FLUX was set to bfloat16 precision because the model was unable to fit in the available memory at higher precision values. This compromise was compensated by the stochastic rounding hyperparameter, which is known to provide better training results for precision levels below 32 bits, such as bfloat16 [49] [42] [38] [37] [40]. The batch size of 4 was chosen to reduce the training time, as for FLUX it was found that smaller batch sizes lead to an increase in training time, but larger batch sizes such as 16 led to significantly worse-looking images. This was because those batch sizes were too large with respect to the size of the data set, leading to convergence on sharp, less generalizable minima within the loss landscape, due to the large batch size causing the optimizer to make big updates to the model parameters. The Adafactor optimizer was found to produce images of better quality and use less memory than the AdamW optimizer after 20 epochs of training with a non-expert visual analysis and was therefore the chosen optimizer for both models.

Both SDXL and FLUX were fine-tuned up to 200 epochs and sampled every 20 epochs using a random seed (including the 0th epoch before fine-tuning) for approximately 200 and 360 fine-tuning hours, respectively, on Nvidia L40 graphical processing units. Subsequently, both models were fine-tuned for an additional 20 epochs

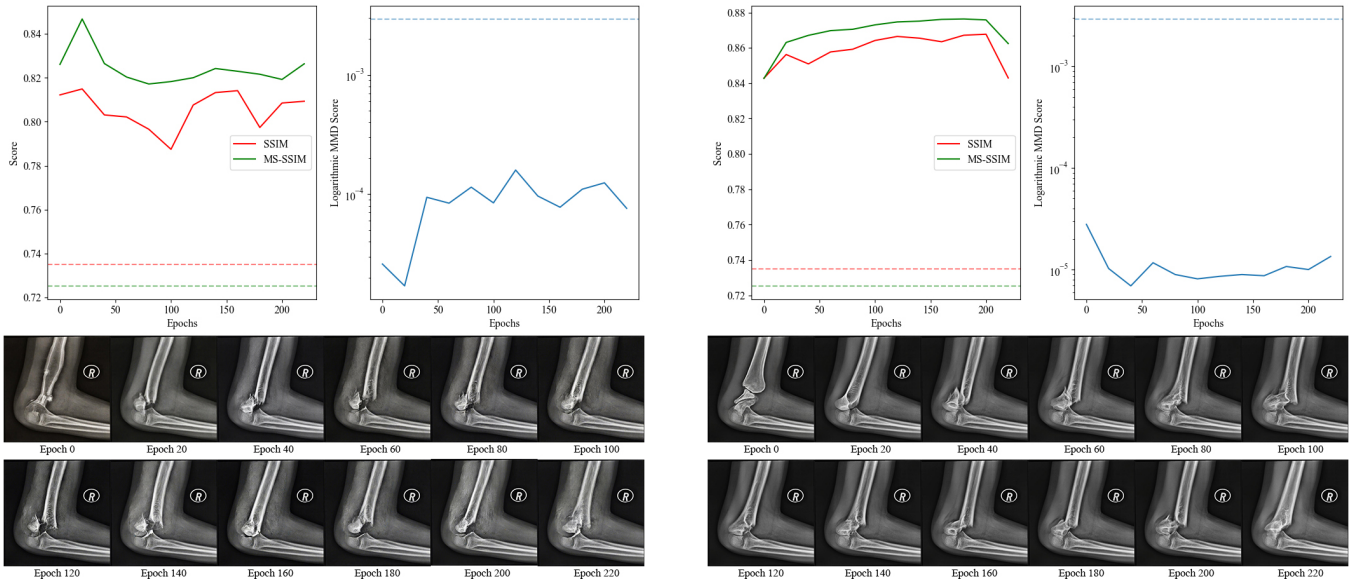

Figure 3: Figures showing the performance of the SDXL (left) and FLUX (right) models. Each figure shows the performance in terms of the SSIM and MS-SSIM scores (upper left) and the logarithmic MMD score using ResNet-50 as the feature extractor (upper right), as the fine-tuning epoch increases. The dashed lines represent the metric scores for images with corresponding masks that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images generated from the one shown in Figure 1 are displayed for each model with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the Epoch 220 results were produced after the model was fine-tuned with a boosted learning rate.

at a boosted learning rate of 1×10^{-4} to determine whether a performance ceiling had been reached.

4 Results

This section provides the main results and discusses some key data points produced. The WGAN-GP, AC-GAN, TAC-GAN, SDXL, and FLUX results are briefly discussed in that order. Interpretations and explanations of these results follow in Section 5. Note that higher scores are better for SSIM and MS-SSIM and lower scores are better for MMD.

4.1 GAN-Based Results

As shown in Figure 4, the WGAN-GP demonstrated steady but slow learning. SSIM and MS-SSIM increased steadily from ≈0.281 to a maximum value of ≈0.292 and ≈0.387 to a maximum value of ≈0.402, respectively, as the training epochs increased. However, MMD showed more unstable learning, which varied between $\approx 2.463 \times 10^{-4}$ and $\approx 3.835 \times 10^{-4}$. From the visual analysis, the WGAN struggled to produce highly realistic images even at the 4000 epoch level. As shown in the sample images displayed in Figure 4, the model learned to create recognizable but completely unrealistic radiographs. Although the overall structure is correctly learned, several issues, such as blurriness, bone merging, unrealistic bone shapes, unrealistic soft tissue, and few recognizable pathologies, are present.

The AC-GAN and TAC-GAN results shown in Figures 5 and Figure 6, respectively, depict the SSIM and MS-SSIM scores immediately decreasing before plateauing as training epochs increase. SSIM for ACGAN decreased from a maximum of ≈0.287 to a minimum of ≈0.261 and MS-SSIM decreased from a maximum of ≈0.407 to a minimum of ≈0.381. SSIM for TAC-GAN decreased from a maximum of ≈0.318 to a minimum of ≈0.299 and MS-SSIM decreased from a maximum of ≈0.426 to a minimum of ≈0.392. Both models showed unstable MMD scores, AC-GAN varied between $\approx 1.373 \times 10^{-4}$ and $\approx 6.369 \times 10^{-4}$ and TAC-GAN varied between $\approx 2.599 \times 10^{-4}$ and $\approx 5.040 \times 10^{-4}$. Visual analysis showed that both AC-GAN and TAC-GAN suffered from similar issues to the WGAN-GP models, with generated images displaying the same realism issues mentioned above. However, of particular note is the generational instability of AC-GAN, as shown by Epoch 3000 in Figure 5, where the generation seed had a large impact on the generated image style and could result in vastly different elbow orientations for separate training epochs. TAC-GAN did not suffer from this and was able to accurately maintain similar elbow orientations between training epochs, as shown in Figure 6.

4.2 Diffusion-Based Results

Upon the visual analysis, SDXL was able to generate images with less blurriness than any of the GAN-based models even at 20 epochs of fine-tuning. Although the unfine-tuned model (Epoch 0) created clear images, noise was present in the images in terms of bone

Metric	Model	Best Score	Noise Improvement
SSIM	WGAN-GP	0.292	3.506
	AC-GAN	0.287	3.439
	TAC-GAN	0.318	3.813
	SDXL	0.815	1.109
	FLUX	0.868	1.180
MS-SSIM	WGAN-GP	0.402	3.965
	AC-GAN	0.407	4.011
	TAC-GAN	0.426	4.204
	SDXL	0.847	1.167
	FLUX	0.876	1.207
MDD	WGAN-GP	2.463×10^{-4}	26.380
	AC-GAN	1.373×10^{-4}	47.311
	TAC-GAN	2.599×10^{-4}	25.001
	SDXL	1.705×10^{-5}	172.168
	FLUX	6.914×10^{-6}	424.483

Table 1: Table showing the respective maximum metric scores and factor of improvement over entirely random noise images (for the GAN-based models) or the random noise-filled masks (for the latent diffusion-based models) for each model. Note that for SSIM and MS-SSIM improvement implies an increase and MMD improvement implies a decrease. All final values were rounded to three decimal places and are thus approximations of the actual observed value.

merging, unrealistic bone shapes, unrealistic soft tissue, few recognizable pathologies, and nonsense items such as unrelated body parts or other unrelated objects being added in place of the mask. Furthermore, it was observed that while fractures were more realistic (less blurry) after 20 epochs of fine-tuning, SDXL struggled with blending the mask area at fine-tuning epochs greater than 20. This was supported by the SSIM and MS-SSIM metrics which, as shown in Figure 3 (Figure 7 shows a full-sized version of this figure), reached their maximum values after 20 epochs of fine-tuning at $\approx$0.815 and $\approx$0.847, respectively. Both metrics saw an immediate decrease after 20 epochs, with SSIM showing notable instability. Furthermore, MMD reached its minimum of $\approx 1.705 \times 10^{-5}$ in 20 epochs of fine-tuning. It should be noted that at all epochs greater than 20, the SSIM, MS-SSIM, and MMD scores were similar or worse than those of even the unfine-tuned model. Another observation is that the boosted learning rate saw worse performance in the visual analysis, SSIM, and MS-SSIM with an insignificant improvement in the MMD score.

In addition to SDXL, FLUX generated visually superior images to the GAN-based models while seeing similar issues with the unfine-tuned model to SDXL. Notably, FLUX saw no blending issues at epochs greater than 20 with image realism improving as the number of fine-tuning epochs increased. This is supported by the SSIM and MS-SSIM metrics, where a constant and steady improvement from $\approx$0.843 to $\approx$0.868 and $\approx$ 0.876, respectively, can be seen in Figure 3 (Figure 8 shows a full-sized version of this figure). The MMD score decreased dramatically to $\approx 6.914 \times 10^{-6}$ at 40 epochs of fine-tuning before plateauing with some variation between $\approx 1.168 \times 10^{-5}$ and $\approx 8.106 \times 10^{-6}$. It should be noted that fine-tuning with the higher learning rate immediately led to worse SSIM, MS-SSIM, and MMD scores.

5 Discussion

This section will explain the reasons for, meaning of, and impacts of the aforementioned results. Initially, the results of the GAN-based methods are discussed and compared with each other, then the results of the latent diffusion-based methods are discussed and compared, and finally the GAN-based and latent diffusion-based results are compared with each other, and the impact of the findings is discussed.

Is it clear from Section 4.1 that WGAN-GP demonstrated the best learning curve. The model showed a steady increase in metric scores as the number of training epochs increased. Unlike AC-GAN and TAC-GAN which showed decreases in their metric scores past 500 training epochs, an indicator that the models were overfitting to specific classes and suffering from mode collapse. This may be explained as the WGAN-GP being able to overcome the issue of mode collapse and ensure training stability through the use of the Wasserstein distance, which provides smoother and more meaningful gradients even for distributions that may not overlap, and the enforcement of the Lipschitz constraint through gradient penalties, which avoids the issues of weight clipping [3] [21]. The introduction of an auxiliary classifier in AC-GAN likely tended to favor generating samples that the classifier found easiest to classify, as observed similarly by Gong et al. [35]. Despite the addition of a second (twin) auxiliary classifier in TAC-GAN, it was observed that TAC-GAN can collapse and struggle with class-conditional diversity similar to AC-GAN, which has been confirmed by previous empirical evidence [2]. However, WGAN-GP did not improve by a large amount in SSIM or MS-SSIM as training epochs increased, with the MMD score demonstrating notable instability and little learning. This may be due to the model being underparameterised or the critic growing too strong, causing the generator to receive weak gradient signals. These findings, along with the fact that the GAN-based models showed a tendency to mode collapse at training resolutions greater than 256x256 pixels, are indicators in support of the research question asked in Section 1.1. That is, it shows that WGAN-GP, AC-GAN, and TAC-GAN may be too unstable when trying to learn patterns in complex adversarial medical data.

As shown in Table 1, TAC-GAN showed the best performance in terms of SSIM and MS-SSIM and the greatest improvement over pure-noise images for these metrics. This means that TAC-GAN produced images that were generally more similar to those found in the test data set than those produced by WGAN-GP or AC-GAN in terms of structural content, contrast, and luminance. Additionally, TAC-GAN showed a 3.813 factor of improvement for SSIM and a 4.204 factor of improvement over pure random noise images, with WGAN-GP and AC-GAN showing similar improvements. This implies that the three models are capable of producing structured images rather than garbage noise and is an indicator that the GAN-based models have learned the underlying patterns in the training

data to some extent. This finding validates the methods performed in Section 3 by showing that the models were able to learn to an extent, which means that they were implemented correctly. The success of AC-GAN over WGAN-GP may be attributed to the addition of the classifier, allowing the model to incorporate more information during generation [53] and the success of TAC-GAN over AC-GAN may be attributed to the additional twin classifier successfully increasing sample class diversity, as has been theoretically shown in previous literature [35]. This finding supports previous literature [53][35], where it was expected that TAC-GAN would outperform AC-GAN and AC-GAN would outperform WGAN-GP. This serves as a further justification that the models were implemented correctly in Section 3. However, the SSIM and MS-SSIM scores for all GAN-based models are low, as shown in previous literature [55]. This means that the models, though outperforming pure noise, generate images that share little structural content with the test data set and fail to preserve perceptually important features at multiple resolutions. Therefore, the produced images are different in both overall structure and fine detail from the real images, further supporting the research question and the hypothesis that these GAN-based models are too simple to accurately model the complex relationships shown in adversarial medical data.

Despite the SSIM and MS-SSIM scores, the MMD score demonstrated more favorable results for the GAN-based models, with AC-GAN scoring the best with 1.373×10^{-4}. This represented nearly a two-time decrease over WGAN-GP and TAC-GAN in the MMD score and means that AC-GAN produced images with a distribution that was closer to the test image distribution than WGAN-GP or TAC-GAN. This contrasts with previous literature [35], which mentioned that TAC-GAN outperformed AC-GAN in terms of MMD when tested on the overlapping MNIST data set; however, this discrepancy may be due to training instability caused by the addition of the twin classifier, as previously reported [31]. This further supports the research question by demonstrating that TAC-GAN may be too unstable during training to create realistic synthetic medical data. All GAN-based models produced good near-zero scores for MMD, which is in contrast to the SSIM and MS-SSIM metrics, where they scored poorly. This is because the SSIM and MS-SSIM metrics are pixel- and structure sensitive, punishing local misalignments, blur, and missing fine detail. While the MMD measures distributional matches in the kernel space, which may be insensitive to local defects. Therefore, these results have shown that although WGAN-GP, AC-GAN, and TAC-GAN can learn and match the overall underlying data distribution, the models are unable to represent fine details and produce non-blurry images. This is visible in Figure 4, Figure 5, and Figure 6, where all three models were able to create rough elbow shapes with correct coloring for the background, soft tissue, and bones, but clearly produce blurry images that lack fine details, such as fracture lines.

From the results reported in Section 4.2 and as shown in Figure 3, FLUX demonstrated a better learning curve than SDXL. This is justified by SDXL reaching optimal scores for SSIM, MS-SSIM, and MMD at 20 epochs of fine-tuning and then plateauing or showing instability. Instead, FLUX shows a smooth and clear improvement as the fine-tuning epochs increase for the SSIM and MS-SSIM metrics; however, the MMD score plateaued after 40 epochs. The better learning curve demonstrated by FLUX is due to the model being larger in complexity and size, incorporating numerous state-of-the-art latent diffusion generation improvements over SDXL. This allowed for greater flexibility of the model to more accurately fit the underlying pattern. It should be noted that this issue may be caused by either a performance ceiling being reached at 20 epochs or by the learning rate being too high for SDXL. Due to the lack of research in the field of SDXL fine-tuning at the time of writing, future research should test even lower learning rates than the one mentioned in Table 7. The plateaued MMD score for FLUX may be explained as the model having reached a near-minimum possible value for the score. The fact that the MMD score did not keep decreasing while the SSIM and MS-SSIM scores continued to increase is an indicator that the model was able to maintain a degree of diversity in the generated samples, rather than overfitting, which would be caused by the distribution of the generated radiographs starting to match the distribution of the real radiographs too closely, causing MMD to continue improving while the SSIM and MS-SSIM metrics would start to worsen. Therefore, this should be viewed as a positive indicator of the model successfully generalizing. It should be noted that when trained to 220 epochs with the boosted learning rate, the model saw a large decrease in the SSIM and MS-SSIM scores and an increase in the MMD score. This is a clear indication that the model was traversing a suitable minimum in the search space, as the higher learning rate caused it to step out of the minimum and produce worse SSIM and MS-SSIM scores. Therefore, a performance ceiling has not yet been reached, and further fine-tuning epochs may have resulted in more realistic images in terms of SSIM and MS-SSIM. These findings support the research question by showing that the FLUX model is better able to learn the underlying patterns in the radiographs than SDXL due to the smoother learning curves demonstrated.

As seen in Table 1, FLUX outperformed SDXL in terms of SSIM, MS-SSIM, and MMD. FLUX also showed greater improvements over images with noise-filled masks, with the MMD score for FLUX being notably better than SDXL. This means that FLUX had greater success in generating radiographs that matched the test radiographs in terms of fine image details, contrast, luminance, and individual and general radiograph structural content, across multiple scales. This is because of both the above mentioned reasons that SDXL is a smaller and simpler model than FLUX and the notable difficulty SDXL experienced in blending the provided radiograph with the filled image mask. As shown in Figure 9, FLUX is able to generate radiographs that have better image blending and demonstrate greater bone, soft tissue, joint, and fracture realism. Despite this, both models showed greater performance than the random noise-filled radiographs, which is an indicator that the models have learned the underlying patterns in the data. The most notable improvement is in the MMD score, where SDXL and FLUX demonstrated factors of improvement of 172.168 and 424.483, respectively. In addition, both models scored highly (scores greater than 0.8) in the SSIM and MS-SSIM metrics, implying that they were able to produce radiographs of high fidelity, low blurriness, high perceptual quality, and radiographs that matched the structural alignment of the test

images. The low, near-zero MMD scores meant that the distributions of the synthetic radiographs were very closely matched to those of the test radiographs. This may be explained in part due to the fact that the latent diffusion-based models only filled the mask space, without affecting the remainder of the radiograph, but also due to the increased complexity of both models, in terms of both the number of parameters and the state-of-the-art latent diffusion improvements incorporated, over previous approaches to diffusion-based synthetic medical image generation [33][1].

Although it is incorrect to directly interpret the SSIM, MS-SSIM, and MMD metrics shown in Table 1 regarding the comparison between the latent diffusion-based and GAN-based models, other implicit metrics such as the performance of each model over random noise and the training curves provide useful insights for comparisons to be made. Although the GAN-based and latent diffusion-based models generated radiographs at different resolutions and the GAN-based models generated entire images rather than filling in a mask like the diffusion models, the shapes of and characteristics demonstrated by the training or fine-tuning plots are independent of these factors. Therefore, it should be said that given the above results, the GAN-based methods demonstrated poor learning due to training instability, mode collapse, and poor generation quality. This is in contrast to SDXL and FLUX which demonstrated good generational ability and particularly FLUX, which demonstrated superior performance to SDXL across all metrics and produced smoother training curves. Due to the differences between mask-based and full-image generation, the noise improvement scores are biased towards the GAN-based models, and the absolute metric scores are biased towards the diffusion-based models. This is because generating an entire image with a competent method should result in a much greater improvement over an entirely random noise image, because even learning the underlying pattern to a small extent will result in large improvements over random noise. This is in opposition to mask-based image generation, where the difference between a competent method accurately filling the mask will be diluted by the remainder of the image, which is shared when noise is only used to fill the mask. Therefore, it was expected that the GAN-based models showed greater noise improvement factors for the SSIM and MS-SSIM metrics; however, it was not expected that SDXL and FLUX demonstrated superior performance with respect to the MMD metric. This indicates that despite the MMD noise improvement factor being biased towards the GAN-based methods, the latent diffusion-based methods were able to leverage the overall image context to generate synthetic radiographs that were orders of magnitude closer to the test radiograph distribution than the GAN-based models. This may be attributed to the better use of information of the masked latent diffusion-based models; incorporating both labels and individual mask context in order to generate radiographs that are focused, very close matches to the real ones, and further emphasizes the likelihood that the GAN-based models suffered from mode collapse. Therefore, this is clear evidence that latent diffusion-based models are better suited for learning the distributions of real medical data.

A key criticism may be raised at this point; that it is unfair to compare the GAN-based methods to the latent diffusion-based methods because the latent diffusion-based methods had to learn and generate smaller areas than the GAN-based methods. However, it should be said that the latent diffusion-based models, while only generating images within the mask area, are, in fact, incorporating more information during the learning process. That is, the latent diffusion-based models were able to reduce the amount of noise learned by focusing specifically on key areas of the radiographs using the image masks. Furthermore, proving whether these latent diffusion-based models are superior to GAN-based methods was not the purpose of this study and has been covered in Section 2. Recall that the purpose of this study was to identify whether masked SDXL and FLUX are better suited for synthetic medical data generation than the unmasked GAN-based methods.

In summary, it has been presented and discussed that the GAN-based models saw several issues when trained on adversarial medical data. Namely; training instability, mode collapse, poor generation resolution, blurriness, and unrealistic radiograph generation. These issues were overcome by the latent diffusion-based models, which demonstrated superior learning curves and improvements over noise for the MMD metric, due to the better use of information and focused generation provided by the image masks. Furthermore, it was shown that of the latent diffusion-based models, FLUX demonstrated higher scores across all metrics and learning curves superior to those of SDXL. These results provide a critical path forward for future research by providing fine-tuning schemes for the state-of-the-art SDXL and FLUX models for the purpose of generating highly realistic synthetic images for medical classification model training.

6 Conclusion

Statistically based machine learning models have been shown to be suitable for the generation of synthetic medical data. Specifically, the FLUX model was identified as superior to SDXL, and both models were shown to be superior to WGAN-GP, AC-GAN, and TAC-GAN in terms of learning the underlying patterns in elbow radiograph images to generate suitably realistic images for training classification models. This has been done by training WGAN-GP, AC-GAN, and TAC-GAN and fine-tuning SDXL and FLUX on a data set of elbow radiographs, evaluating the generated synthetic radiographs in terms of the SSIM, MS-SSIM, and MMD metrics, and then discussing and interpreting these metrics alongside analyzing the training curves of each model. The GAN-based models were identified to have several issues that make them unsuitable for realistic synthetic medical image generation, and the latent diffusion-based models were able to leverage state-of-the-art improvements over the GAN-based models to generate highly realistic synthetic elbow radiographs. This study represents a technical contribution to the field of synthetic medical data generation by providing a clear path for future research. That is, future research should focus on training a classification model on the synthetic data produced, the impact that data mixing has on CAD model classification accuracy, determining optimal fine-tuning hyperparameters for SDXL and FLUX, and incorporating an analysis by medical experts to gauge the realism of synthetic radiographs in a real-world setting.

7 Acknowledgments

The author thanks Oliver Foxcroft for developing the metrics used for model comparisons, Ethan Topat for developing and training the GAN-based models, Nicholas Kruger for providing the data set, and Geoff Nitschke and Bilal Aslan for their supervision.

References

[1] Abdullah al Nomaan Nafi et al. 2024. Diffusion-Based Approaches in Medical Image Generation and Analysis. arXiv:2412.16860 [eess.IV] https://arxiv.org/abs/2412.16860

[2] Valérie Mezger Anis Bourou and Auguste Genovesio. 2024. GANs Conditioning Methods: A Survey. *arXiv preprint arXiv:2408.15640* (2024).

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

[4] Ali Borji. 2019. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding* 179 (Feb. 2019), 41–65. doi:10.1016/j.cviu.2018.10.009

[5] Axel Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer.

[6] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG] https://arxiv.org/abs/2105.05233

[7] Kunaal Dhawan and Siddharth Nijhawan. 2024. Cross-modality synthetic data augmentation using gans: Enhancing brain mri and chest x-ray classification. *MedRxiv* (2024), 2024–06.

[8] Arthur Gretton et al. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13, 25 (2012), 723–773. http://jmlr.org/papers/v13/gretton12a.html

[9] Arash Marioriyad et al. 2024. Diffusion Beats Autoregressive: An Evaluation of Compositional Generation in Text-to-Image Models. arXiv:2410.22775 [cs.CV] https://arxiv.org/abs/2410.22775

[10] Abhinau Venkataramanan et al. 2021. A hitchhiker's guide to structural similarity. *IEEE Access* 9 (2021), 28872–28896.

[11] Bilal Aslan et al. 2025. Deep-Learning Classifiers for Small Data Orthopedic Radiology. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*. IEEE, 1–7.

[12] Burçin Kurt et al. 2012. Medical images enhancement by using anisotropic filter and CLAHE. In *2012 International symposium on innovations in intelligent systems and applications*. IEEE, 1–4.

[13] Changhee Han et al. 2018. GAN-based synthetic brain MR image generation. In *IEEE: International symposium on biomedical imaging*. Washington DC, USA, 734–738. doi:10.1109/ISBI.2018.8363678

[14] Chenglin Yang et al. 2024. 1.58-bit FLUX. arXiv:2412.18653 [cs.CV] https://arxiv.org/abs/2412.18653

[15] Dara Hallinan et al. 2021. International transfers of personal data for health research following Schrems II: a problem in need of a solution. *European journal of human genetics* 29, 10 (2021), 1502–1509.

[16] Dustin Podell et al. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952

[17] Fabi Prezja et al. 2022. DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Scientific Reports* 12, 1 (Nov 2022). doi:10.1038/s41598-022-23081-4

[18] Gao Huang et al. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV] https://arxiv.org/abs/1608.06993

[19] Hanxi Sun et al. 2023. A deep learning approach to private data sharing of medical images using conditional generative adversarial networks (Gans). https://pmc.ncbi.nlm.nih.gov/articles/PMC10325103/

[20] Illya Bakurov et al. 2022. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Systems with Applications* 189 (2022), 116087.

[21] Ishaan Gulrajani et al. 2017. Improved Training of Wasserstein GANs. arXiv:1704.00028 [cs.LG] https://arxiv.org/abs/1704.00028

[22] James Betker et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

[23] Jack Hessel et al. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV] https://arxiv.org/abs/2104.08718

[24] Jason Itri et al. 2018. Fundamentals of diagnostic error in imaging. *Radiographics* 38, 6 (2018), 1845–1865.

[25] Junnan Li et al. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[26] Jianlin Su et al. 2023. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv:2104.09864 [cs.CL] https://arxiv.org/abs/2104.09864

[27] Jascha Sohl-Dickstein et al. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585 [cs.LG] https://arxiv.org/abs/1503.03585

[28] Ke Gu et al. 2015. Quality assessment considering viewing distance and image resolution. *IEEE Transactions on Broadcasting* 61, 3 (2015), 520–531.

[29] Kaiming He et al. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[30] Kaiyi Huang et al. 2023. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 78723–78747. https://proceedings.neurips.cc/paper_files/paper/2023/file/f8ad010cdd9143dbb0e9308c093aff24-Paper-Datasets_and_Benchmarks.pdf

[31] Ligong Han et al. 2020. Unbiased auxiliary classifier gans with mine. *arXiv preprint arXiv:2006.07567* (2020).

[32] Luis More et al. 2015. Parameter tuning of CLAHE based on multi-objective optimization to achieve different contrast levels in medical images. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4644–4648.

[33] Loc Nguyen et al. 2023. A New Chapter for Medical Image Generation: The Stable Diffusion Method. In *2023 International Conference on Information Networking (ICOIN)*. 483–486. doi:10.1109/ICOIN56518.2023.10049010

[34] Mostafa Dehghani et al. 2023. Scaling Vision Transformers to 22 Billion Parameters. arXiv:2302.05442 [cs.CV] https://arxiv.org/abs/2302.05442

[35] Mingming Gong et al. 2019. Twin Auxilary Classifiers GAN. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/4ea06fbc83cdd0a06020c35d50e1e89a-Paper.pdf

[36] Mira Hayati et al. 2023. Impact of CLAHE-based image enhancement for diabetic retinopathy classification through deep learning. *Procedia Computer Science* 216 (2023), 57–66.

[37] Naveen Mellempudi et al. 2019. Mixed precision training with 8-bit floating point. *arXiv preprint arXiv:1905.12334* (2019).

[38] Naigang Wang et al. 2018. Training deep neural networks with 8-bit floating point numbers. *Advances in neural information processing systems* 31 (2018).

[39] Patrick Esser et al. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV] https://arxiv.org/abs/2403.03206

[40] Pedram Zamirai et al. 2020. Revisiting bfloat16 training. *arXiv preprint arXiv:2010.06192* (2020).

[41] Robin Rombach et al. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[42] Suyog Gupta et al. 2015. Deep learning with limited numerical precision. In *International conference on machine learning*. PMLR, 1737–1746.

[43] Wei Wang et al. 2021. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems* 34, 1 (2021), 264–277.

[44] Yaron Lipman et al. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747 [cs.LG] https://arxiv.org/abs/2210.02747

[45] Zhou Wang et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[46] Centers for Disease Control and Prevention et al. 2003. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. *MMWR: Morbidity and mortality weekly report* 52, Suppl 1 (2003), 1–17.

[47] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. 2023. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. arXiv:2310.11513 [cs.CV] https://arxiv.org/abs/2310.11513

[48] Kamal Halloum and Hamid Ez-Zahraouy. 2025. Enhancing Medical Image Classification through Transfer Learning and CLAHE Optimization. *Current Medical Imaging* 21, 1 (2025), e15734056342623.

[49] Markus Höhfeld and Scott Fahlman. 1992. Probabilistic rounding in neural network learning with limited precision. *Neurocomputing* 4, 6 (1992), 291–299.

[50] Sun ju Byeon et al. 2022. Automated histological classification for digital pathology images of colonoscopy specimen via deep learning. *Scientific Reports* 12, 1 (2022), 12804.

[51] Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

[52] Nerogar. 2025. OneTrainer. https://github.com/Nerogar/OneTrainer.

[53] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 2642–2651. https://proceedings.mlr.press/v70/odena17a.html

[54] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4195–4205.

[55] Shubham Rana and Matteo Gatti. 2025. Comparative Evaluation of Modified Wasserstein GAN-GP and State-of-the-Art GAN Models for Synthesizing Agricultural Weed Images in RGB and Infrared Domain. *MethodsX* 14 (2025), 103309.

[56] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. 2019. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. *Journal of Computer and Communications* 7, 3 (March 2019), 8–18. doi:10.4236/jcc.2019.73002 Number: 3 Publisher: Scientific Research Publishing.

[57] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV] https://arxiv.org/abs/1409.1556

[58] Isaac Wagner and Kaustubh Chakradeo. 2025. Human-AI Complementarity in Diagnostic Radiology: The Case of Double Reading. *Philosophy & Technology* 38, 2 (2025), 1–31.

[59] Zhou Wang, Eero Simoncelli, and Alan Bovik. 2003. Multiscale structural similarity for image quality assessment. In *IEEE: Asilomar conference on signals, systems and computers*, Vol. 2. California, USA, 1398–1402. doi:10.1109/ACSSC.2003.1292216

[60] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2022. Simple Multi-Dataset Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7571–7580.

A Appendix

A.1 Table of Experiments

Number	Name	Methodology	Outcome
1	Human non-expert image evaluation	A sample of 50 synthetic images per model at specific epoch intervals was manually reviewed by a non-expert to identify obvious error areas such as extreme blurriness.	The GAN-based models suffered from blurriness and other realism issues, SDXL struggled with mask blending, FLUX generated good quality images.
2	SSIM Metric Test	Compute and plot the SSIM metric at specific training and fine-tuning intervals	TAC-GAN produced the highest score for the GAN-based models and FLUX produced the highest score for the latent diffusion-based models
3	MS-SSIM Metric Test	Compute and plot the MS-SSIM metric at specific training and fine-tuning intervals	TAC-GAN produced the highest score for the GAN-based models and FLUX produced the highest score for the latent diffusion-based models
4	MMD Metric Test	Compute and plot the MMD metric at specific training and fine-tuning intervals	AC-GAN produced the lowest score for the GAN-based models and FLUX produced the lowest score for the latent diffusion-based models

Table 2: Table of experiments showing the experiments performed to address the research question. Note that only one non-expert evaluated the images due to resource limitations.

A.2 Data Set Distributions

Pathology	Train Count (% Total)	Test Count (% Total)
Distal humerus fracture	60 (2.83%)	25 (2.78%)
Elbow dislocation anterior	12 (0.57%)	4 (0.45%)
Elbow dislocation posterior	36 (1.70%)	15 (1.67%)
Joint effusion	432 (20.39%)	184 (20.49%)
Lateral epicondyle displaced	99 (4.67%)	42 (4.68%)
Medial epicondyle displaced	66 (3.12%)	27 (3.01%)
Olecranon fracture	45 (2.12%)	19 (2.12%)
Proximal radial fracture	34 (1.61%)	14 (1.56%)
Proximal ulnar metaphysis fracture	24 (1.13%)	9 (1.00%)
Radial head fracture	17 (0.80%)	7 (0.78%)
Radial head subluxation	17 (0.80%)	6 (0.67%)
Soft tissue swelling	215 (10.15%)	91 (10.13%)
Supracondylar fracture	609 (28.74%)	261 (29.07%)
Normal	453 (21.38%)	194 (21.60%)
Total	**2119**	**898**

Table 3: A table showing the data set breakdown for the training and test data sets by pathology. There are 14 classes in total.

A.3 WGAN-GP Architecture

Operation	Kernel	Strides	Feature Maps	BN	Dropout	Nonlinearity
G(z) - 256 x 1 x 1 input						
Linear	N/A	N/A	512 x 16 x 16	✓	0.0	ELU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	256 x 32 x 32	✓	0.0	ELU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	128 x 64 x 64	✓	0.0	ELU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	64 x 128 x 128	✓	0.0	ELU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	32 x 256 x 256	✓	0.0	ELU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	1 x 256 x 256	×	0.0	Tanh
D(x) - 1 x 256 x 256 input						
Conv2D	5 x 5	2 x 2	64 x 128 x 128	×	0.25	ELU(α=0.2)
Conv2D	5 x 5	2 x 2	128 x 64 x 64	×	0.25	ELU(α=0.2)
Conv2D	5 x 5	2 x 2	256 x 32 x 32	×	0.25	ELU(α=0.2)
Conv2D	5 x 5	2 x 2	512 x 16 x 16	×	0.25	ELU(α=0.2)
Conv2D	5 x 5	2 x 2	512 x 8 x 8	×	0.25	ELU(α=0.2)
Linear	N/A	N/A	1	×	0.0	N/A
Optimizer	Adam (α=0.0002, β_1=0.5, β_2=0.9, weight_decay=0.0001)					
Batch Size	5					
Epochs	4000					

Table 4: Table showing the WGAN-GP Generator (G) and Discriminator (D) architectures and training hyperparameters.

A.4 AC-GAN Architecture

Operation	Kernel	Strides	Feature Maps	BN	Dropout	Nonlinearity
G(z, y) - 256 x 1 x 1 + 18 x 1 x 1						
Linear	N/A	N/A	256 x 4 x 4	✓	0.0	N/A
Upsample + Conv2D	3 x 3	1 x 1	128 x 8 x 8	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	64 x 16 x 16	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	32 x 32 x 32	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	16 x 64 x 64	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	8 x 128 x 128	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	4 x 256 x 256	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	1 x 256 x 256	×	0.0	Tanh
D(x) - 1 x 256 x 256 input						
Conv2D	3 x 3	2 x 2	4 x 128 x 128	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	8 x 64 x 64	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	16 x 32 x 32	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	64 x 8 x 8	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	128 x 4 x 4	✓	0.25	LeakyReLU(α=0.2)
Linear	N/A	N/A	1	×	0.0	N/A
Linear	N/A	N/A	18	×	0.0	N/A
Generator Optimizer	Adam (α=0.0005, β_1=0.5, β_2=0.999)					
Discriminator Optimizer	Adam (α=0.0002, β_1=0.5, β_2=0.999)					
Classifier Optimizer	Adam (α=0.0002, β_1=0.5, β_2=0.999)					
Batch Size	5					
Epochs	4000					

Table 5: Table showing the AC-GAN Generator (G) and Discriminator (D) architectures and training hyperparameters.

A.5 TAC-GAN Architecture

Operation	Kernel	Strides	Feature Maps	BN	Dropout	Nonlinearity
G(z, y) - 256 x 1 x 1 + 18 x 1 x 1						
Linear	N/A	N/A	256 x 4 x 4	✓	0.0	N/A
Upsample + Conv2D	3 x 3	1 x 1	128 x 8 x 8	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	64 x 16 x 16	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	32 x 32 x 32	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	16 x 64 x 64	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	8 x 128 x 128	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	4 x 256 x 256	✓	0.0	LeakyReLU(α=0.2)
Upsample + Conv2D	3 x 3	1 x 1	1 x 256 x 256	×	0.0	Tanh
D(x) - 1 x 256 x 256 input						
Conv2D	3 x 3	2 x 2	4 x 128 x 128	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	8 x 64 x 64	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	16 x 32 x 32	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	64 x 8 x 8	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	128 x 4 x 4	✓	0.25	LeakyReLU(α=0.2)
Linear	N/A	N/A	1	×	0.0	N/A
Linear	N/A	N/A	18	×	0.0	N/A
C(x) - 1 x 256 x 256 input						
Conv2D	3 x 3	2 x 2	4 x 128 x 128	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	8 x 64 x 64	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	16 x 32 x 32	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	64 x 8 x 8	✓	0.25	LeakyReLU(α=0.2)
Conv2D	3 x 3	2 x 2	128 x 4 x 4	✓	0.25	LeakyReLU(α=0.2)
Linear	N/A	N/A	18	×	0.0	N/A
Generator Optimizer	Adam (α=0.0005, β_1=0.5, β_2=0.999)					
Discriminator Optimizer	Adam (α=0.0002, β_1=0.5, β_2=0.999)					
Classifier Optimizer	Adam (α=0.0002, β_1=0.5, β_2=0.999)					
Batch Size	5					
Epochs	4000					

Table 6: Table showing the TAC-GAN Generator (G), Discriminator (D) and Classifier (C) architectures and training hyperparameters.

A.6 SDXL and FLUX Hyperparameters

Hyperparameter	SDXL	FLUX
Weight data type	float32	bfloat16
Prior data type	float32	bfloat16
Text encoder data type	float32	bfloat16
VAE data type	float32	float32
Train data type	float32	bfloat16
Output data type	float32	bfloat16
Output resolution	1024x1024	1024x1024
Epochs	200	200
Learning rate	1×10^{-5}	1×10^{-5}
Learning rate scheduler	Constant	Constant
Batch size	4	4
Optimizer	Adafactor	Adafactor
Scale parameter	False	False
Relative step size	False	False
Warm-up initialization	False	False
Stochastic rounding	False	True
Fused back pass	False	False
Decay rate	-0.8	-0.8
Text encoder learning rate	3×10^{-6}	3×10^{-6}
Text encoder training epochs	200	200
EMA	False	False
Train transformer (UNET)	True	True
Transformer training epochs	200	200
Unmasked probability	0.1	0.1
Unmasked weight	0.4	0.4
Normalize area loss	False	False
Loss weight function	Constant	Constant
Gamma	5.0	5.0
Loss scaler	None	None

Table 7: Table showing the chosen hyperparameters for the SDXL and FLUX models.

A.7 WGAN-GP Results

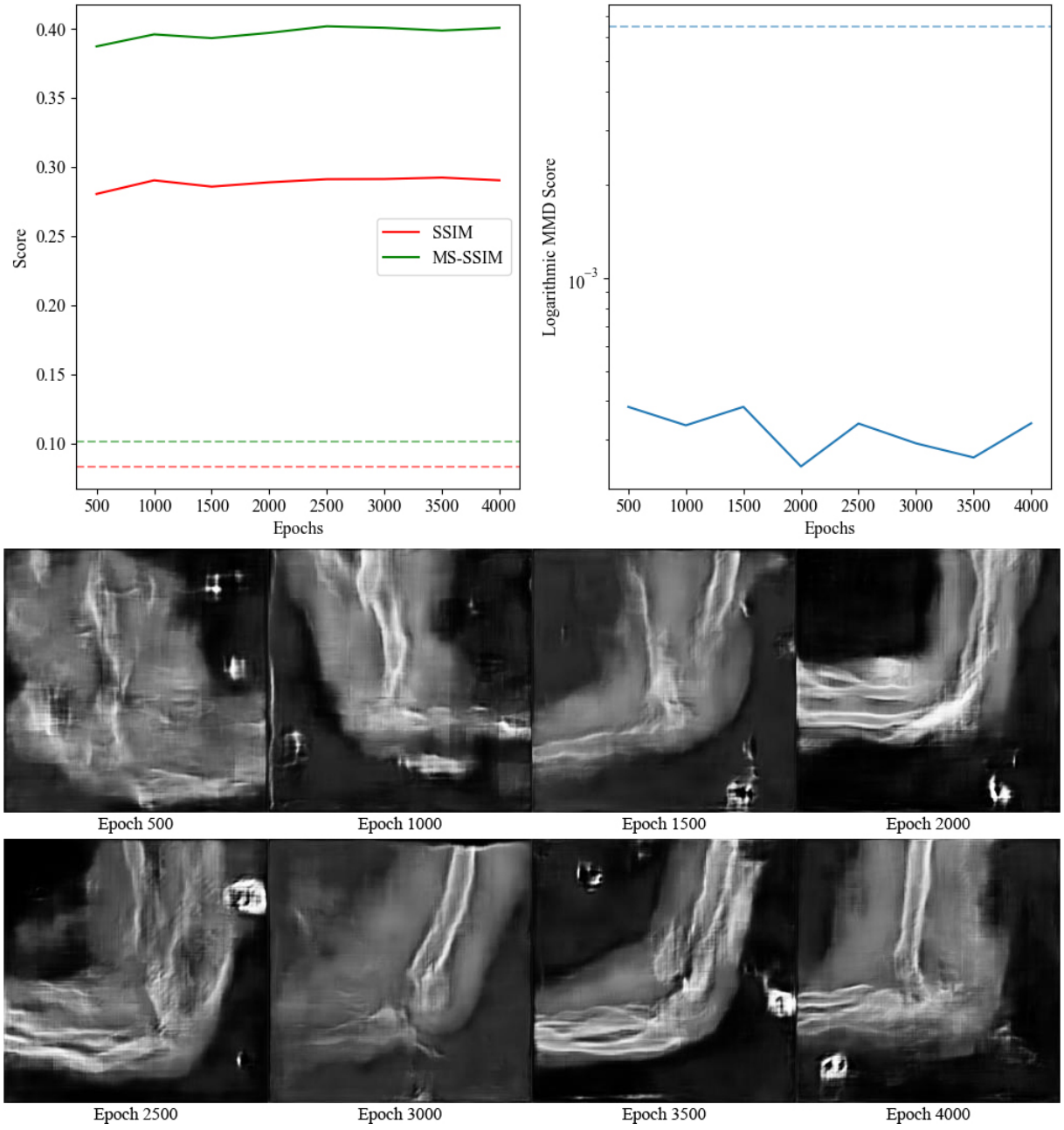

Figure 4: Figure showing the performance of the WGAN-GP model in terms of the SSIM and MS-SSIM scores (left) and the logarithmic MMD score using ResNet-50 as the feature extractor (right), as the training epoch increases. The dashed lines represent the metric scores for images that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images shown are attempting to replicate the real radiograph shown in Figure 1 and are displayed with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the WGAN-GP did not support labels or masks and the image was generated using the LAT supracondylar fracture trained model.

A.8 AC-GAN Results

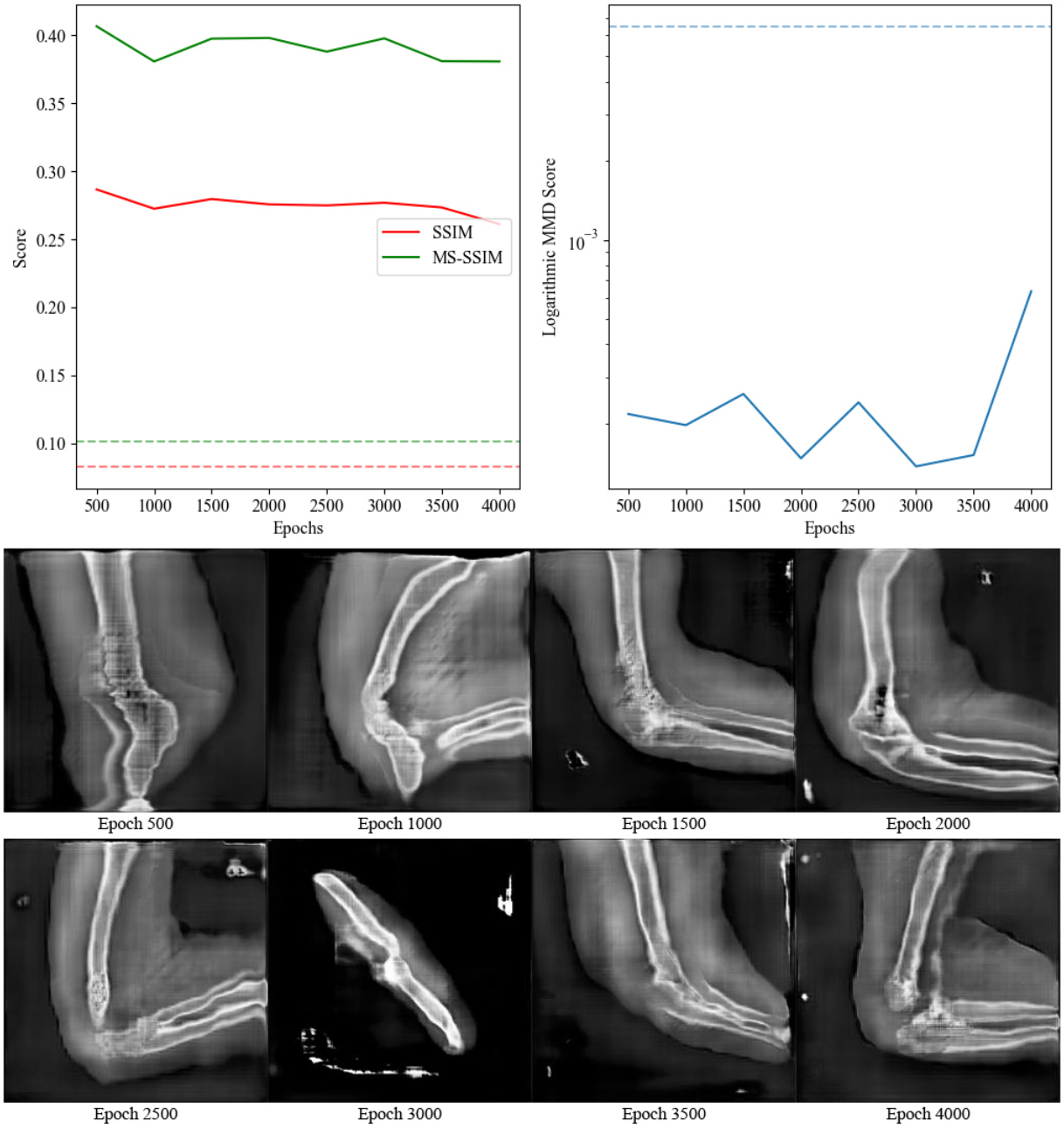

Figure 5: Figure showing the performance of the AC-GAN model in terms of the SSIM and MS-SSIM scores (left) and the logarithmic MMD score using ResNet-50 as the feature extractor (right), as the training epoch increases. The dashed lines represent the metric scores for images that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images shown are attempting to replicate the real radiograph shown in Figure 1 and are displayed with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the AC-GAN did not support masks and the image was generated using only the provided one-hot encoded label.

A.9 TAC-GAN Results

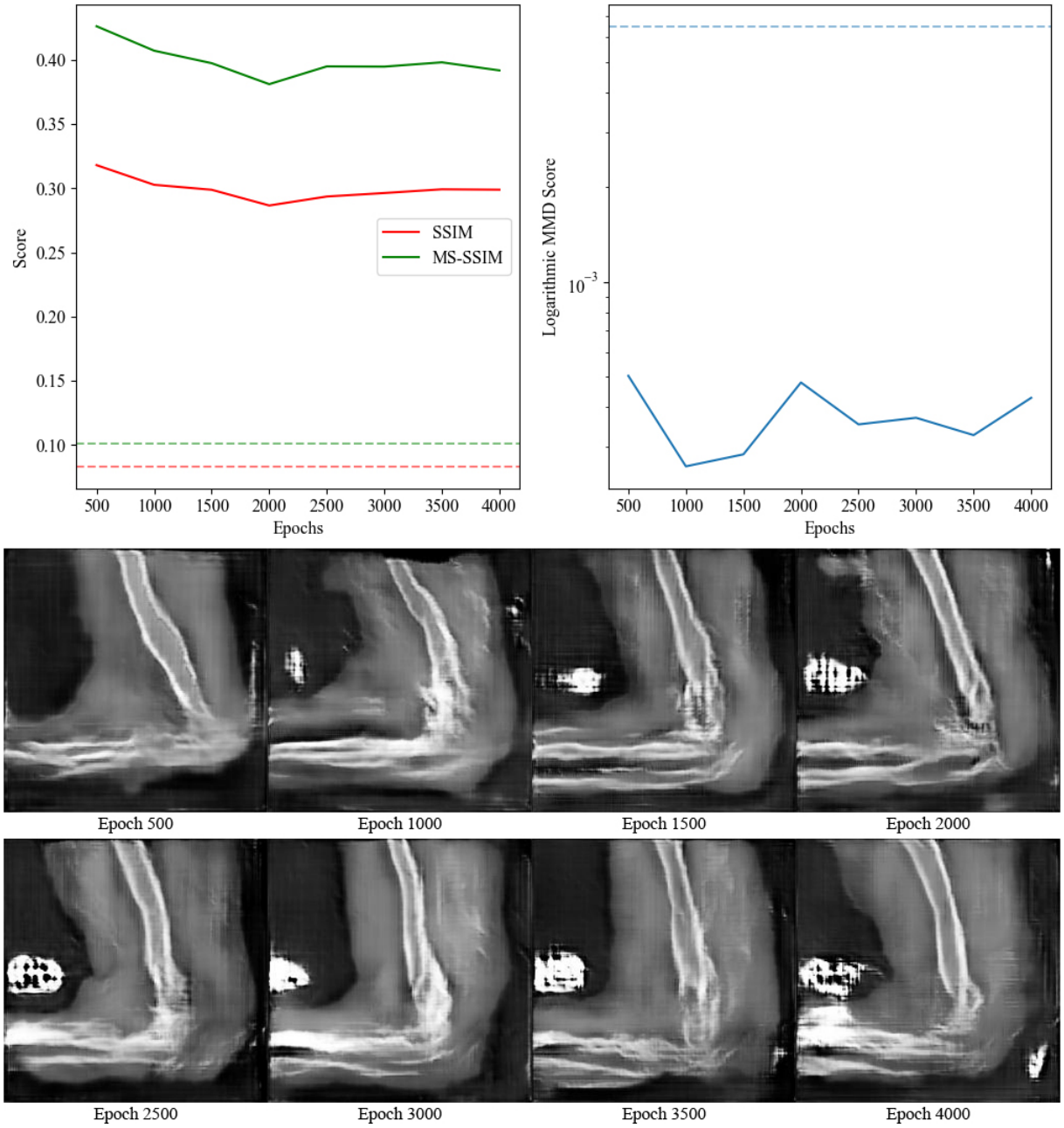

Figure 6: Figure showing the performance of the TAC-GAN model in terms of the SSIM and MS-SSIM scores (left) and the logarithmic MMD score using ResNet-50 as the feature extractor (right), as the training epoch increases. The dashed lines represent the metric scores for images that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images shown are attempting to replicate the real radiograph shown in Figure 1 and are displayed with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the TAC-GAN did not support masks and the image was generated using only the provided one-hot encoded label.

A.10 SDXL Results

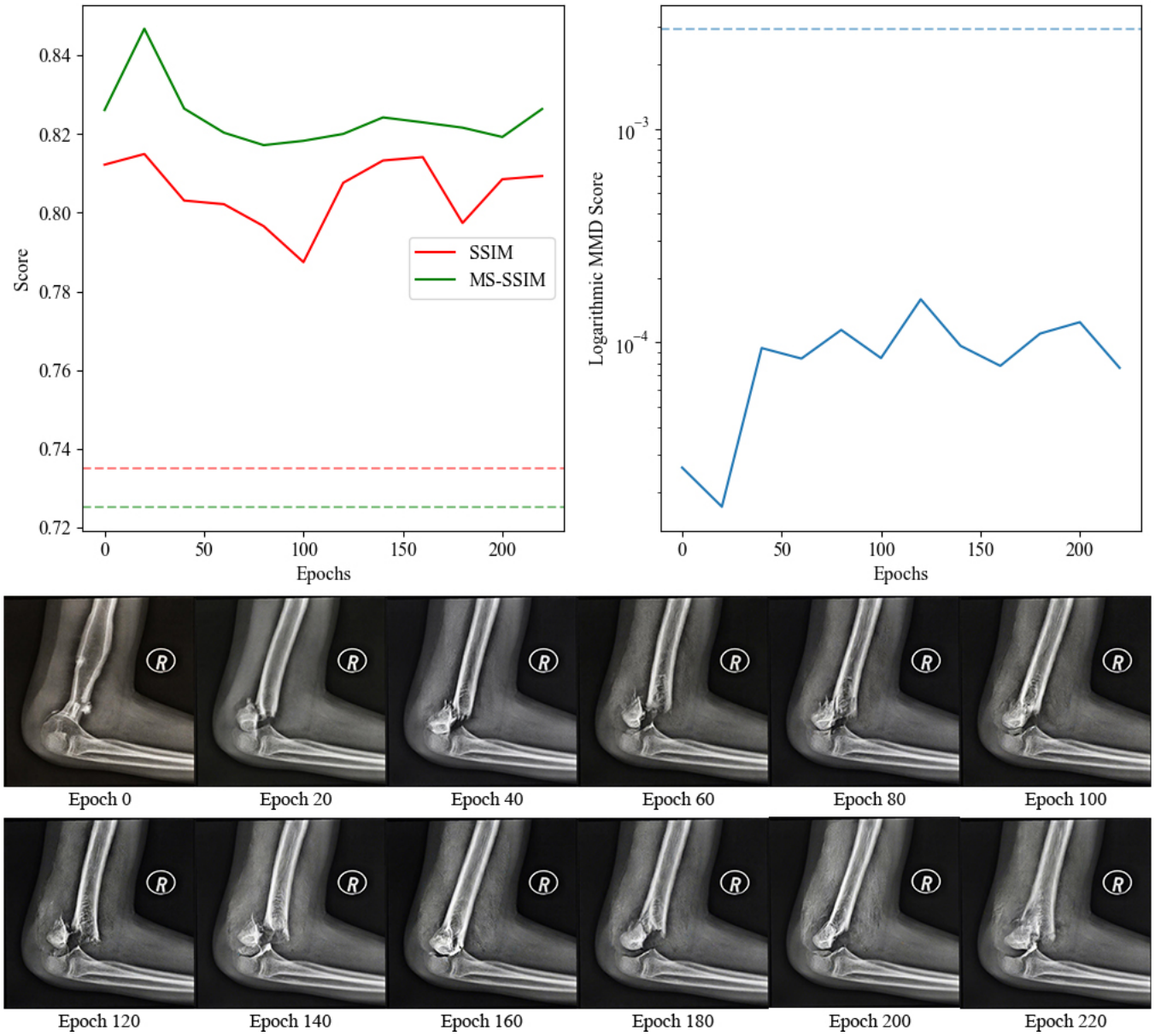

Figure 7: Full-size figure showing the performance of the SDXL model in terms of the SSIM and MS-SSIM scores (left) and the logarithmic MMD score using ResNet-50 as the feature extractor (right), as the fine-tuning epoch increases. The dashed lines represent the metric scores for images with corresponding masks that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images generated from the one shown in Figure 1 are displayed with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the Epoch 220 results were produced after the model was fine-tuned with a boosted learning rate.

A.11 FLUX Results

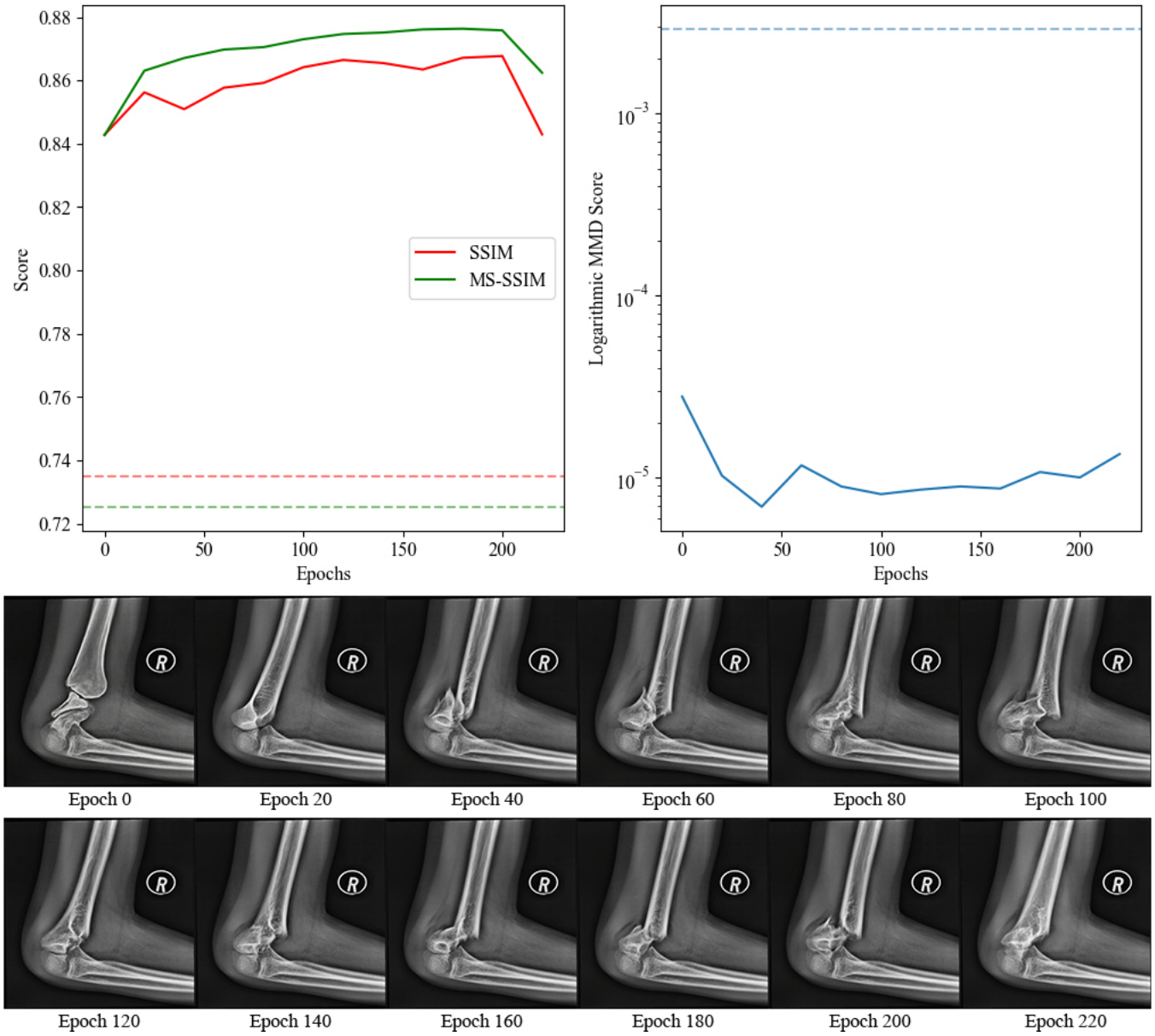

Figure 8: Full-size figure showing the performance of the FLUX model in terms of the SSIM and MS-SSIM scores (left) and the logarithmic MMD score using ResNet-50 as the feature extractor (right), as the fine-tuning epoch increases. The dashed lines represent the metric scores for images with corresponding masks that were entirely made up of random noise, forming baseline minimum expected scores for the corresponding metric. The sample images generated from the one shown in Figure 1 are displayed with the fine-tuning epoch increasing from left to right and from top to bottom (below). Note that the Epoch 220 results were produced after the model was fine-tuned with a boosted learning rate.

A.12 SDXL Comparison To FLUX

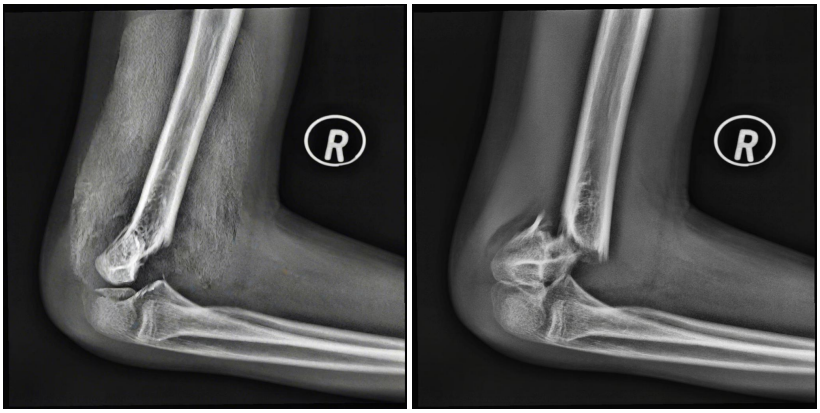

Figure 9: Figure showing the synthetic images generated at epoch 200 by SDXL (left) and FLUX (right). Both images were generated from the sample image shown in Figure 1. Several visual differences are present and show that FLUX produces visually higher quality images than SDXL at higher fine-tuning epochs. Namely; synthetic image blending, bone realism, soft tissue realism, joint realism, and fracture realism.