Innovation/Impact: Generative adversarial networks (GANs) have recently been used in medical imaging for anomaly detection, data augmentation for computer-aided diagnosis, image translation, image segmentation, treatment planning, and image reconstruction. Past works in natural imaging have demonstrated that limited data leads to poor-quality images and that image quality can be improved using transfer learning and data augmentation. As most medical imaging datasets are of limited size, understanding how transfer learning and data augmentation affect the performance of StyleGAN2 on a medical dataset is of utmost importance. We are the first to apply StyleGAN2 to a high-resolution medical imaging dataset. Additionally, we provide empirical evidence that the Fréchet Inception Distance (FID) is consistent with human perceptual evaluation of medical images.

Key Results: Transfer learning, data augmentation, and including additional data points resulted in substantial improvements upon the baseline (see Table 1). The best results were achieved when transfer learning and data augmentation were used together. In fact, they resulted in a greater improvement (63%) of the FID than increasing the size of the dataset about fifteenfold (44%). Example synthetic images from Experiments 1, 4, and 8 are shown in Figure 1. Many of the images generated by Experiment 1 contain noise artifacts, especially in the liver. The images generated by both Experiments 4 and 8 demonstrate reduced noise artifacts, enhanced detail, and superior anatomical accuracy.

Experiment	FID
1. 10,600 images	15.18
2. 10,600 images w/ pretraining	11.22
3. 10,600 images w/ data augmentation	7.24
4. 10,600 images w/ pretraining & data augmentation	**5.65**
5. 20,579 images	9.71
6. 44,578 images	10.79
7. 84,799 images	9.50
8. 153,945 images	8.51

Table 1: The quantitative results of our eight experiments. The reported FIDs are the best FIDs achieved in 10,000 training iterations.

Through four visual Turing tests (VTT) given to six participants with a medical imaging background, we confirmed that transfer learning and data augmentation improved the perceptual quality of the images (see Table 2). When both methods were used in tandem, participants were more likely to say a synthetic image was real than fake. As the FID is a metric that was created for natural images, recent works have argued that it is not applicable to medical images. In contrast, our results show that the FID is consistent with human perceptual judgement on medical images: the lower the FID, the higher the average false positive rate (Pearson correlation of -0.91, 90% confidence).

Experiment	FID	FPR (%)	FNR (%)
1. 10,600 images	10.43	29 (±27)	32 (±21)
2. 10,600 images w/ pretraining	7.78	34 (±19)	32 (±18)
3. 10,600 images w/ data augmentation	7.15	49 (±11)	34 (±18)
4. 10,600 images w/ pretraining & data augmentation	**5.06**	**55 (±9)**	**41 (±11)**

Table 2: The average false positive rates (FPRs) and false negative rates (FNRs) acquired from the multi-model VTT. The reported FIDs are the best FIDs achieved in 25,000 training iterations. The model weights associated with these best FIDs were used to create the synthetic images for the VTT.

The results of a VTT given to seven clinicians (shown in Table 3) confirm the high-quality nature of the synthetic images from Experiment 4. Overall, the clinicians classified synthetic images as real 42% of the time, approaching the equivalent of random guessing. As part of the VTT, evaluators were given a Likert scale evaluating the realness of each image. Real images achieved an average rating of 3.99 (±1.00) and synthetic images an average rating of 3.23 (1.21).

	Precision (%)	Recall (%)	Accuracy (%)	FPR (%)	FNR (%)
Average	66 (±12)	72 (±21)	65 (±10)	42 (±27)	28 (±21)

Table 3: The results of the VTT given to radiologists and radiation oncologists. The generative model used in this test had a FID of 5.06 (Experiment 4, 25k training iterations).

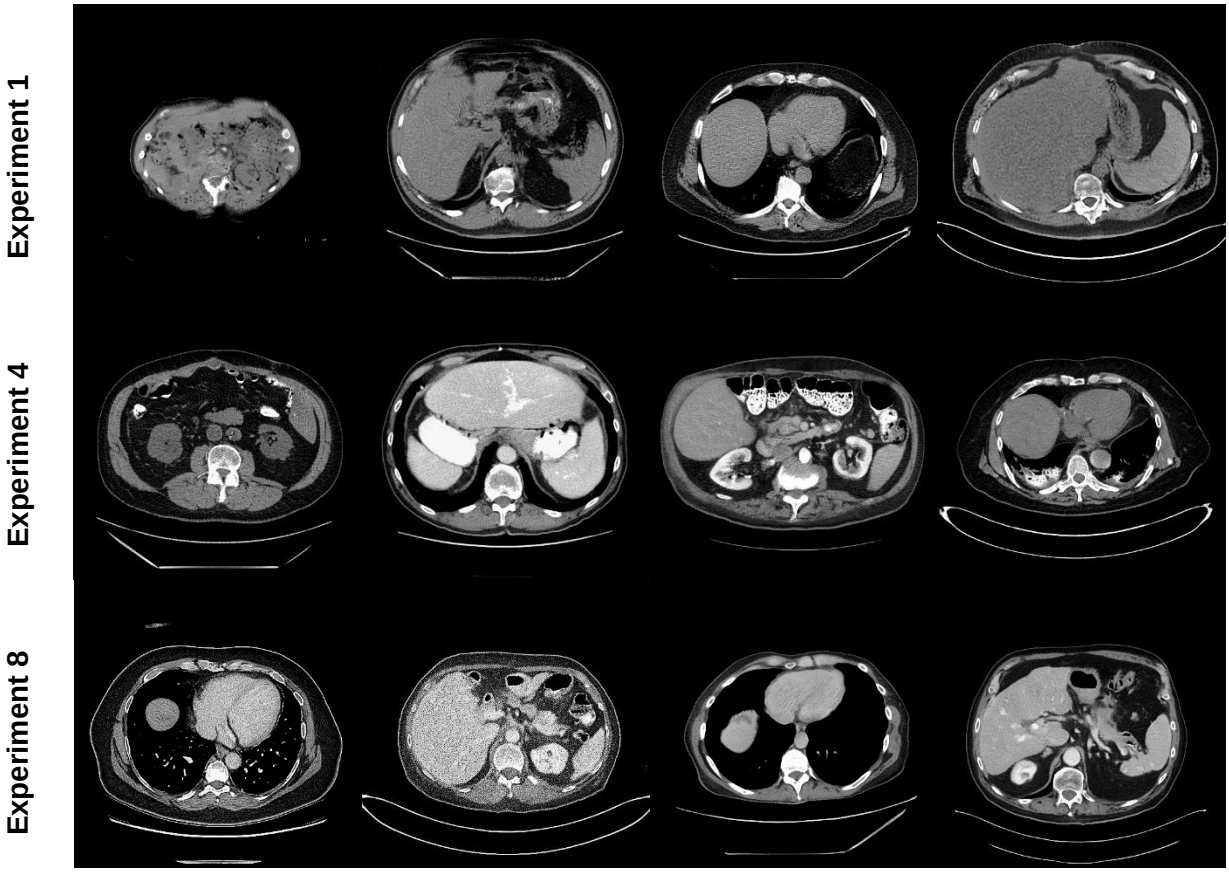

Figure 1: Example generated images from Experiments 1, 4, and 8. The first row contains generated images from Experiment 1 (10k images). The second row contains generated images from Experiment 4 (10k images with pretraining and data augmentation). The third row contains generated images from Experiment 8 (155k images). All images are generated from the model weights associated with the reported FID scores in Table 1. All images were randomly selected.

One reason that pretraining and augmentation on a small dataset outperformed baseline training on a larger dataset is due to the discriminator overfitting on the training samples. When the discriminator overfits, the generative and discriminative losses diverge. In Figure 2, we see that for all experiments where augmentation was not used (Experiments 1-2, 5-8), the losses diverged before 500 ticks. Adding 143,345 images from 301 new patients was not enough to eliminate this training divergence, whereas augmentation successfully stabilized training.

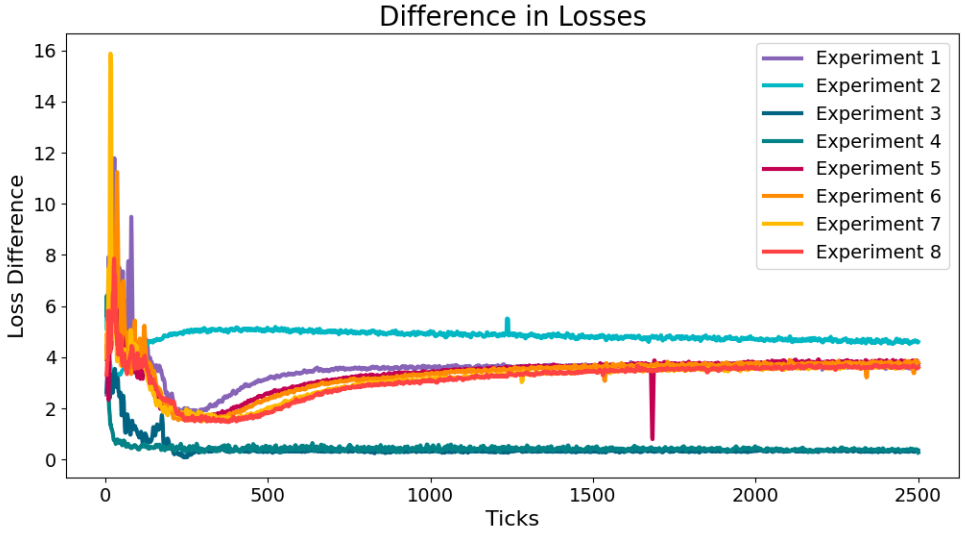

Figure 2: The simple moving average of the absolute difference between the generative and discriminative losses over training.