Anthony Vo

CS598 – Deep Learning for Healthcare (Spring 2025)

7 May 2025

Final Project Report – Automated Treatment Planning in Radiation Therapy using GANs

Research Paper

Mahmood, R., Babier, A., McNiven, A., Diamant, A. & Chan, T.C.Y.. (2018). Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks. Proceedings of the 3rd Machine Learning for Healthcare Conference, in Proceedings of Machine Learning Research 85:484-499 Available from https://proceedings.mlr.press/v85/mahmood18a.html.

Dataset

Craft, D., Bangert, M., Long, T., Papp, D., & Unkelbach, J. (2014). Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset. Gigascience, 3(1). https://doi.org/10.1186/2047-217x-3-37

Video Link

https://mediaspace.illinois.edu/media/t/1_dw0w78va

GitHub Repo Link

https://github.com/voacado/AutoTreatmentPlanningRadiationTherapyGANs/tree/main

Kaggle (Training) Link

https://www.kaggle.com/code/voanthony/automated-treatment-planning-using-gans

Introduction

In the research paper "Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks", the authors address one of the primary bottlenecks in oncologic care: the labor-intensive process that is generating clinically acceptable dosage distributions for radiotherapy patients. Mahmood et al. (2018) propose a Generative Adversarial Network (GAN) framework that learns a direct mapping from patient CT images to three-dimensional dose predictions (as another form of imagery). Thus, bypassing the need for the iterative decision process traditionally required by radiotherapy treatment planners. Their core contribution is the fact that a data-driven, adversarial model can produce high-quality dose plans that closely match those generated by expert planners. It can do so in mere seconds, a significant reduction in planning time compared to the hours it would take prior, paving the way for more scalable and consistent radiotherapy treatment workflows.

The wider research space in machine learning for healthcare has seen generative models applied for various tasks such as image generation, segmentation, and prognostic modeling. However, their use in treatment planning remains under-developed. By leveraging both the adversarial loss to encourage human-like realism as well as reconstruction loss to ensure the model's performance stays true to professional-grade doctor's intuitions, the model proposed in this paper attempts to bridge the gap between rapid-fire data-driven inference and the extremely strict clinical requirements for these life-and-death high-accuracy clinical scenarios. This paper not only validates the feasibility of GANs in this domain, but also establishes various quantitative benchmarks such as dose-volume histogram (DVH) metrics for comparing AI-generated plans versus conventional optimization methods for radiotherapy.

In terms of scope of reproducibility, we were able to re-produce a significant amount of the paper. While we do not have access to the dataset they used directly to train their GAN model, we had access to the CORT (Common optimization dataset for radiation therapy) dataset which is extremely similar. This dataset contains intensity modulation radiation therapy (IMRT) datasets for a prostate case, liver case, and a head and neck case. For the sake of scalability (since I am training this GAN model on Kaggle's Nvidia P100 GPU), I chose to stick with just the liver case as it contained a sufficient size dataset (while not being too extremely large) to showcase that this methodology works.

For paper reproduction, we were able to recreate the dataset ingestion and processing pipeline, with dose volumes extracted out of the datasets as well as their CT scan photos from the MAT files. We were able to replicate the model architecture here as well, since the research paper utilizes a pix2pix architecture as its foundation with a GAN attached. Here, we reimplemented the pix2pix autoencoder architecture and utilized a PatchGAN architecture to recreate the generator and discriminator networks in PyTorch to similar success. We additionally compared the generated dose distributions against those published as baselines and observed similar results (we will talk more about this in the "Results" section below).

If I had additional time, I would have loved to attempt this project on the head-and-neck dataset of CORT, which is about 100-times larger.

Methodology

Our development environment was based on Python 3.11.11 running on Ubuntu 22.04.4 LTS with a Nvidia P100 and Intel Xeon CPU (6th gen @ 2Ghz). Core dependencies include: numpy == 1.26.4, scipy == 1.15.2, pydicom == 3.0.1, matplotlib==3.7.5, scikit-image==0.25.1, torch==2.5.1+cu124. The entire training environment was handled by the free version of Kaggle on the P100 accelerator.

We utilize the liver case radiotherapy data from the CORT (Common optimization dataset for radiation therapy) dataset. This consists of co-registered CT volumes and clinically approved 3D dose maps in MATLAB format. After downloading the .mat and .dcm files from their public website (http://gigadb.org/dataset/100110 – required files: "LIVER.zip" and "Liver_dicom.zip" on Page 7 in Files – or my extract on Kaggle: https://www.kaggle.com/datasets/voanthony/imrt-optimization-research-the-cort-dataset), we write pre-processing code to resample our data to a uniform voxel grid, normalize Hounsfield units to [-1,+1], create 3D masks for the data structures based on "_VOILIST.mat" files, and create synthetic dose maps based on the provided PTV (planning target volume) and OAR (organs at risk) data. In the form of a DataLoader for PyTorch, this is an example of one of the data elements:

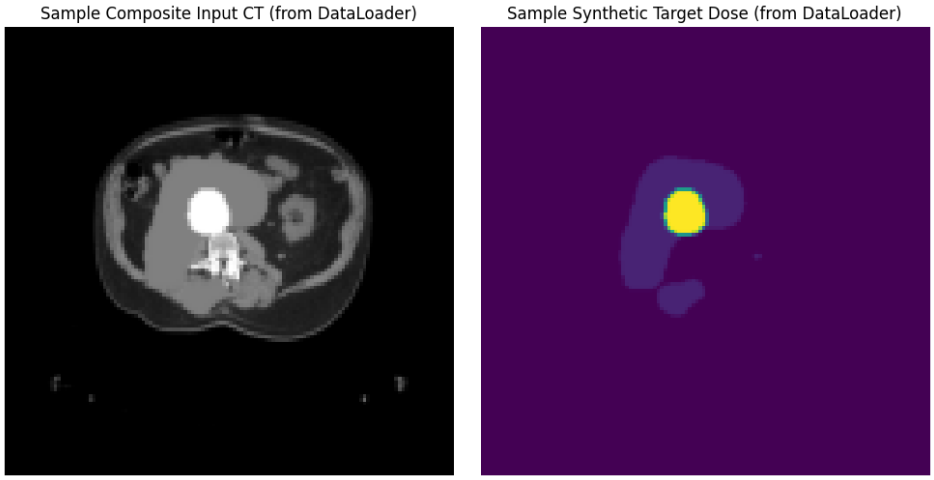


Figure 1 - Example data (from the liver case in the CORT dataset)

The model architecture is based on the Mahmood et al.'s proposed architecture found at this repo: https://github.com/rafidrm/gancer/tree/master. The generator employs an encoder-decoder structure with skip connections, minimizing the joint loss:

$$\min_G \max_D \quad Ex, y[\log D\,(x, y)] + Ex\left[\log\left(1 - D\big(x, G(x)\big)\right)\right] + \lambda|y - G(x)|_1$$

where x is the CT input, y is the ground-truth dose, and λ balances the adversarial and L1 terms. The discriminator is a PatchGAN that classifies local image patches to encourage high-frequency detail. Rather than a normal GAN, PatchGAN's patches methodology allows it to capture more detailed information about the image in smaller chunks, which better represents how a radiotherapy doctor may approach their dose optimization (observing only portions of the CT that are of interest rather than wasting time thinking about the deadspace).

The training methodology involved 25 epochs with a learning rate of $2 * 10^{-4}$, a batch size of 4, and an Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$). Dropout of 0.5 was applied in the decoder. Hardware comprised of a Nvidia P100 (16GB) with an average runtime of 0.5 seconds per epoch (roughly 2 minutes of compute time to train). Training combined the adversarial loss with the regularization (L1) loss as described above.

To evalulate the results, we utilized Mean-Absolute Error (MAE) and Peak Signal-to-Noise Ratio (PSNR) methods. Mean Absolute Error captures the overall agreement between predicted and ground-truth doses, and its equation is described as follows:

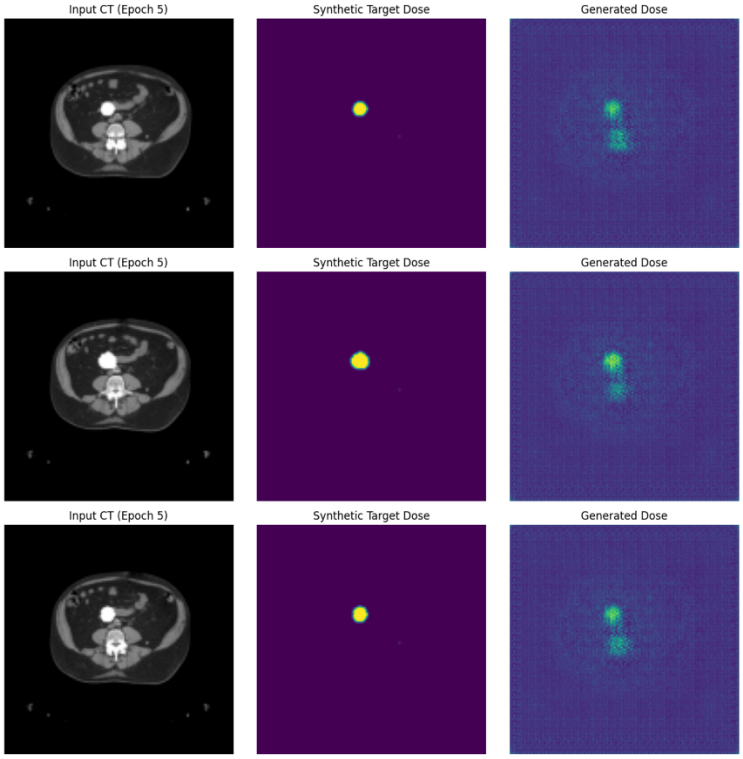
$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}(\, d_i^{\text{pred}} - d_i^{\text{true}})^2$$

For Peak Signal-to-Noise Ratio (PSNR), it measures noise (in decibels) to quantify the similarity between two images. It does so by measuring the ratio between the maximum possible signal power and power of the noise (error) that affects the fidelity of its representation. Thus, a higher PSNR value (in decibels) indicate that the predicted and true dose distributions are more similar (lower "noise" or error). Typically medical-image comparisons range from 20 dB (low quality) to over 40 dB (high quality). PSNR is defined as follows:

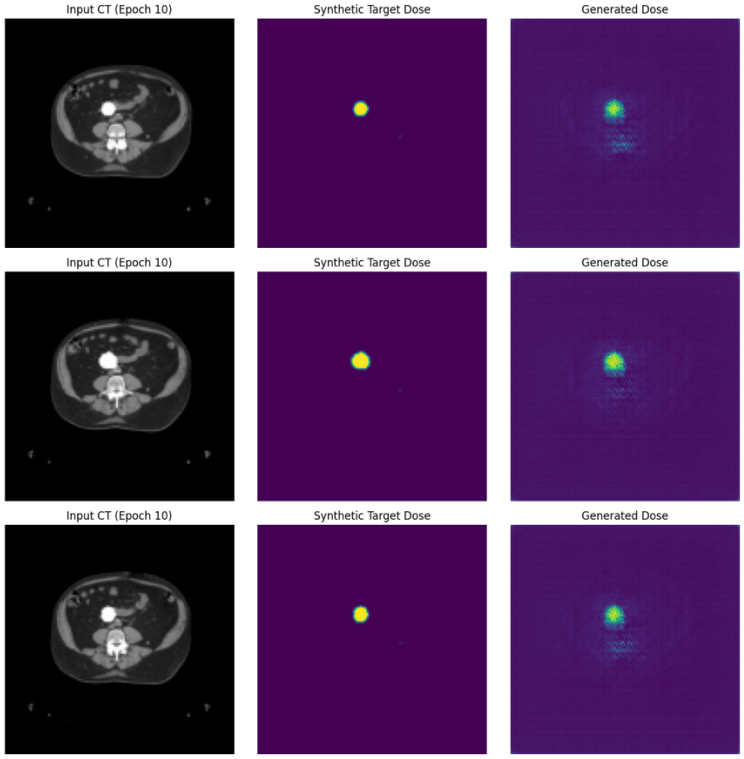
$$PSNR = 20log_{10}\left(\frac{D_{max}}{\sqrt{MSE}}\right)$$

Results

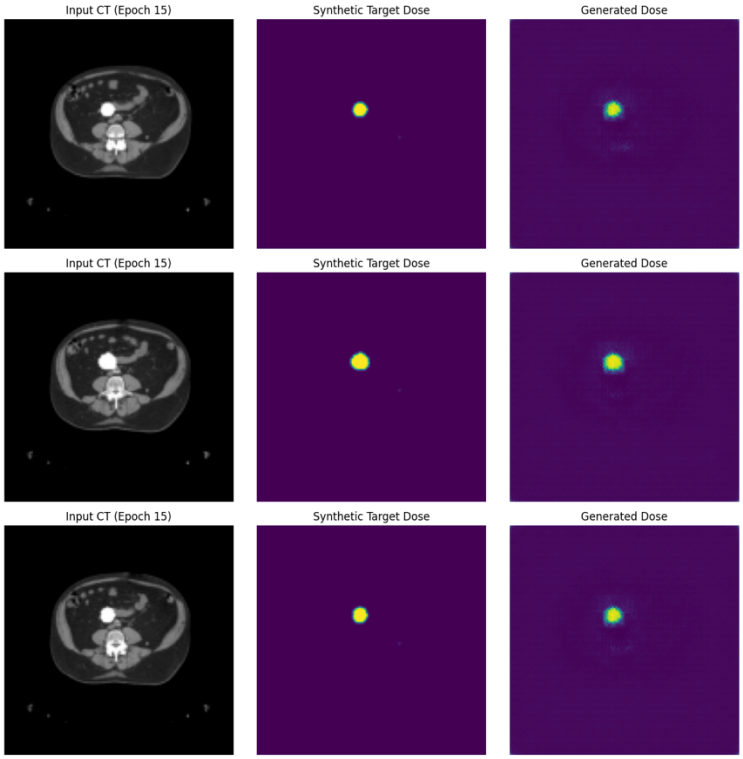
We can observe the model performance at various epochs:



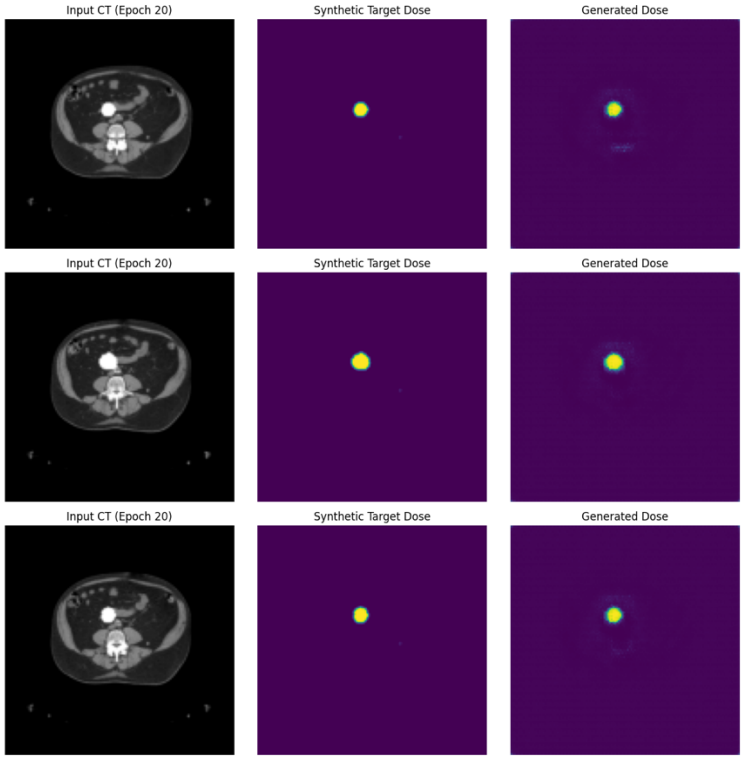
Epoch 5



Epoch 10

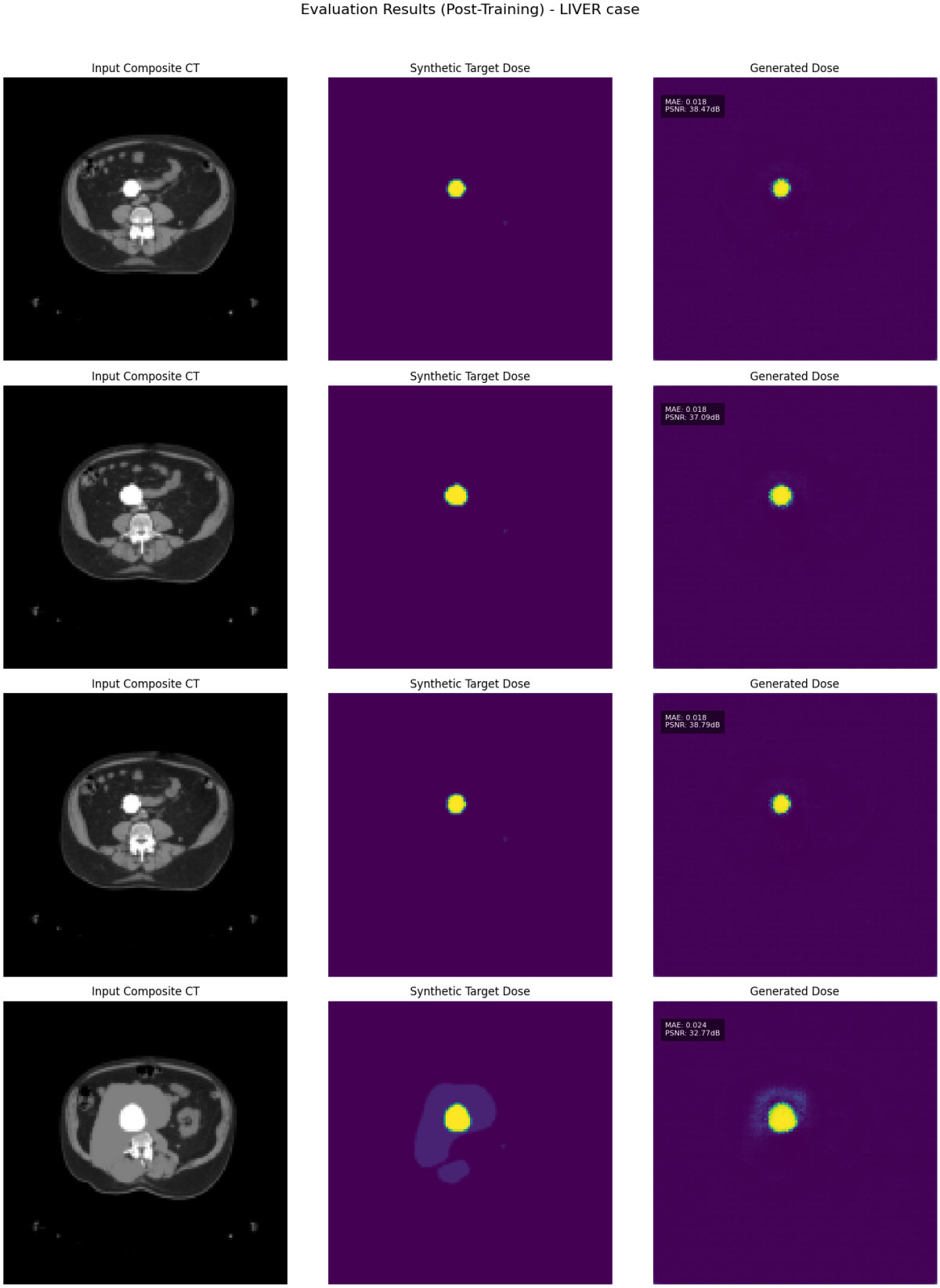


Epoch 15

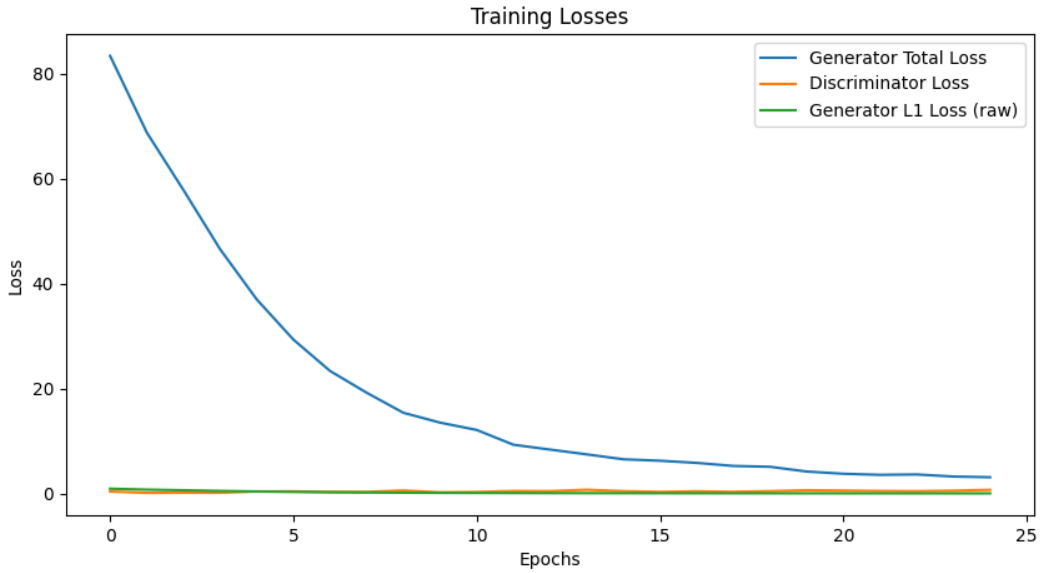


Epoch 20

At epoch 25:


Evaluation Results (Post-Training) - LIVER case
Input Composite CT
Synthetic Target Dose
Generated Dose
MAE: 0.018
PSNR: 38.47dB
Input Composite CT
Synthetic Target Dose
Generated Dose
MAE: 0.018
PSNR: 37.09dB
Input Composite CT
Synthetic Target Dose
Generated Dose
MAE: 0.018
PSNR: 38.79dB
Input Composite CT
Synthetic Target Dose
Generated Dose
MAE: 0.024
PSNR: 32.77dB

And in summary:



Here, we can see that, when replicating the methodology verbatim, the results are very strong. Looking at the two validation metrics we have defined, Mean-Absolute Error and Peak Signal-to-Noise Ratio – our five validation test cases performed extremely well with very low MAE (average of 0.0195) and higher PSNR (average of 36.78dB, which lands in the upper-percentile of PSNR outputs).

Compared to the original Mahmood et al. (2018) study, our liver-dataset results show up a roughly 1.95% relative error versus the roughly 3.6% relative error in the head-and-neck case (2.15 Gy out of a 60 Gy prescription) in the paper, and about a 4.3 dB higher PSNR. Thus, our higher PSNR suggests that our GAN is producing liver dose maps with finer fidelity to the ground truth (clinical research) than what was achieved on the head-and-neck data. However, these may not be direct comparisons as the liver dataset is about $1/100^{th}$ as large as the head-and-neck data, so there may be far more noise in the dataset Mahmood et al. choose to use.

There are a few extensions that would be worthwhile to implement into this project. For example, the most obvious addition would be extending this to organ parts of the body outside just the head-and-neck, liver, and prostate. What about your breast cancer, etc.? The one that I chose to implement was PSNR validation, which was not originally used in the paper. This is an alternative metric we can use to validate our model performance, and provides an interpretable scalar that complements the validation metrics used in the paper: DVH and γ-analysis. This additional validation metric seems viable, as a lower MAE seems generally correlated with a higher PSNR dB, which makes sense given how it compares between two images.

Discussion

The experimental results on the liver dataset: sub-2% average MAE and PSNR in the high 30 dB range, emphasize the GAN's capacity to learn robust dose mappings across on the liver dataset. Given the results of Mahmood et al. on head-and-neck data, it highlights the GAN's capacity to learn dose mappings across different anatomies, preserving dose fidelity but also capturing high-frequency dose variations critical for clinical usage.

This original paper is extremely reproducible due to its simpler nature, easy to replicate data source, and their official GitHub replace which contains the full generator and discriminator code and training scripts as inspiration. The largest setback in replicating this paper was mostly in creating the DataLoaders to work with this data architecture, but that was a short-lived problem. Additionally, piecing together the dataset as various parts were in different files (such as the previously mentioned "_VOILIST.mat" files).

For the authors, I would recommend they containerize their project and environment into a Dockerfile or Conda environment with the exact package versions and system requirements so that spin-up is much faster. I would also appreciate more effort in suggesting similar data sources – while the CORT dataset did eventually work, I would like to see implementation logic using it.

Author Contributions

As this is a solo project, the entirety of the project was completed by Anthony Vo.