Novel Lung CT Image Synthesis at Full Hounsfield Range with Expert Guided Visual Turing Test

Arjun Krishna, Shanmukha Yenneti, Ge Wang and Klaus Mueller

May 16, 2023

Novel Lung CT Image Synthesis at Full Hounsfield Range With Expert Guided Visual Turing Test

Arjun Krishna[1], Shanmukha Yenneti[1], Ge Wang[2], and Klaus Mueller[1]

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA
[2]Biomedical Imaging Center, School of Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract Conventional image quality metrics are unsuitable to evaluate the realism and medical accuracy of synthetically generated CT images. We describe an approach based on the concept of Visual Turing Test that engages medical professionals to assess the generated images and provide useful feedback that can inform the generative process. We first describe our approach for synthesizing large numbers of novel and diverse CT images across the full Hounsfield range using a very small annotated dataset of around thirty patients and a large non-annotated dataset with high resolution medical images. Using an anatomy exploration interface we can generate CT images with anatomies that were non-existent within either of the datasets, without compromising accuracy and quality. Our approach works for all Hounsfield windows with minimal depreciation in anatomical plausibility. We then describe our Visual Turing Test methodology in detail and show results we have obtained.

1 Introduction

Deep learning in medical applications is limited due to the low availability of large labeled, annotated or segmented training datasets. The scarcity persists not only because of privacy and ownership concerns but also because of the high cost of labeling such datasets by human experts. Likewise, publicly available annotated high resolution image datasets are also often very small or even non-existent.

In this work we first present a methodology that reduces or even eliminates he problem of such small datasets by converting them into large datasets without the loss of anatomical accuracy. Our approach goes beyond simple data augmentation techniques like stretching or flipping existing images and adds new data instances with anatomies that may not even exist in these datasets. With this approach we are able to increase not only the size but the overall diversity of images in datasets significantly.

Our method uses a dataset of segmented CT images from thirty patients and a large dataset of unsegmented CT images. Our method builds on our previous work of texture learning [1] to expand the small annotated dataset with textures present in the large dataset. Subsequently we extract segmentation maps from the unsegmented large dataset via a trained U-Net. Next we train a cycleGAN on both the small segmented data and large unsegmented data in an alternate fashion to generate new images with segmentation maps as inputs. This synthesis step expands on our previous work [2] and explores the PCA space of segmentation maps in conjunction with the cycleGAN to create CT images with novel anatomies not present in either of the datasets.

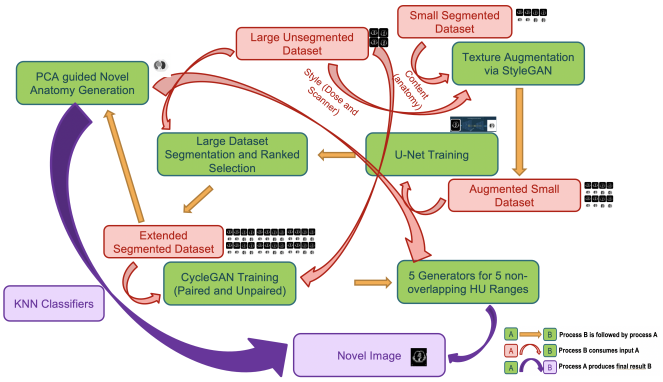

Small Segmented Dataset
Large Unsegmented Dataset
Texture Augmentation via StyleGAN
PCA guided Novel Anatomy Generation
Style (Dose and Scanner)
Content (anatomy)
Large Dataset Segmentation and Ranked Selection
U-Net Training
Augmented Small Dataset
Extended Segmented Dataset
CycleGAN Training (Paired and Unpaired)
5 Generators for 5 non-overlapping HU Ranges
KNN Classifiers
Novel Image
A B Process B is followed by process A
A B Process B consumes input A
A B Process A produces final result B

Figure 1: Flow starts at the top right corner with two datasets - a small segmented and a large unsegmented dataset. Three different Deep-Learning networks are used starting from a StyleGAN followed by a U-NET segmentation network and 5 CycleGANs which train generators for the final step.

Since commonly used image quality metrics are unsuitable to evaluate the realism and medical accuracy of synthetically generated CT images, we have designed a framework that engages medical professionals to assess the generated images along these qualitative figures of merit. Our evaluation interface is based on the concept of Visual Turing Test and provides several design elements to determine the degree of realism and the sources of anatomical imperfections.

2 Our CT Synthesis Methodology

Figure 1 highlights our sequence of steps. We will briefly summarize each step in the same sequence below.

Texture Augmentation. The smaller dataset consists of chest CT scans with segmentation maps (lungs, heart, etc.) of 30 patients. The larger dataset consists of non-annotated chest CT scans of ∼14k patients. To use the two datasets together we modified the textures of the smaller dataset with those of the larger one, augmenting the smaller annotated dataset 3-fold. We used the network architecture of [1] for segment-wise texture learning and created new CT images with the anatomy from the small dataset and the textures from the larger dataset.

Further Augmentation from Label Training. We train a U-Net [3] to output a segmentation map given a chest CT image as input. We use the augmented annotated dataset created in the previous step for training our U-Net. Having similar textures across the two different datasets helps in training a segmentation network on one dataset to segment the images of another. We use the trained U-Net to segment

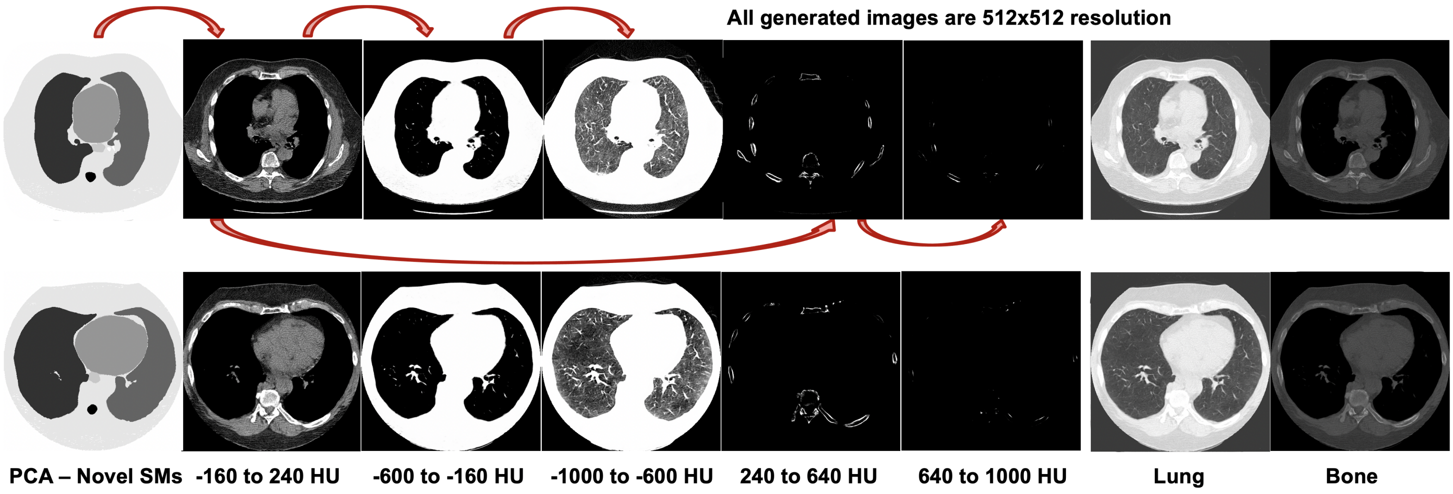


Figure 2: Above figure shows two examples of novel CT scans generations. The sequential training and generation learns the correlations of anatomical details and can be clearly seen within the columns as we move from left to right. The last two columns depict the anatomical consistency observed in different HU windows than in generated ones. Each red arrow represents a generator of the two generators trained in a cycleGAN setup for corresponding modalities.

all 14k patient images. Since the smaller dataset has limited anatomy, there are errors in the segmentation outputs of the larger dataset. k-NN classifiers are used to rank them by accuracy using certain characteristics of the segmentation images. We choose the best 1/4 of segmentation outputs and add them with their CT scans to the smaller segmented dataset. This dataset along with the larger dataset of unsegmented images is then used to train the generators for the synthesis.

Decomposing the Hounsfield Range for Generation Steps. Our method generates images at full Hounsfield in five separate steps. Fig. 3a shows the average distribution of pixels values of a chest CT-scan over HU values. Fig. 3b shows an image in (-160, 240) HU range while Fig. 3c shows an image in (-600, -1000) HU ranges. Two separate generators are used to generate these HU ranges thereby assisting the GANs to focus on the minute details within these ranges since discriminators within a GAN setup focus on the accuracy of the majority group of pixels within a particular HU range. Hence we use five generators to generate five distinct sets of images for five distinct HU ranges for a single CT image generation. We first generate the middle HU range image using the segmentation map as input since it details the major anatomical features such as bones and organs. We then use this generated image as input for generating the other HU range images. This is shown in detail in Fig. 2.

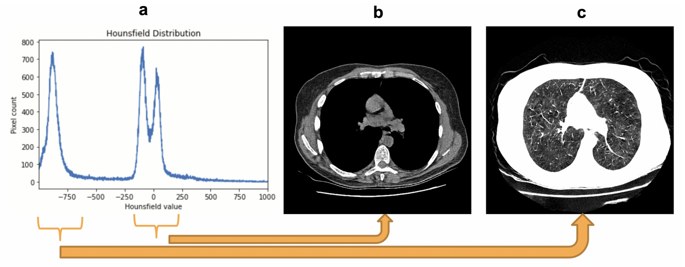


Figure 3: We use 5 CycleGANs to train 5 generators for 5 non-overlapping HU ranges (-1000, -600), (-600, -160), (-160, 240), (240, 640), (640, 1000)

Paired and Unpaired Training via CycleGAN. To generate the CT images we follow the network architecture of [4] for paired and unpaired training. We use a different algorithm and data setup for training since our paired and unpaired datasets come from different sources. We use only the large CT dataset for unpaired training while we use all the segmentation maps for both paired and unpaired training. Training was done in an alternate fashion; every iteration of paired training was followed by two iterations of unpaired training to learn the anatomical diversity present in the unsegmented dataset. As mentioned before, we have five such setups to produce five relevant generators to cover all five HU ranges. Figure 2 shows the image synthesis sequence we use to cover the full HU-range. Shown are two CT images which exhibit novel anatomy. The left two columns demonstrate their anatomical consistency in the lung and bone windows.

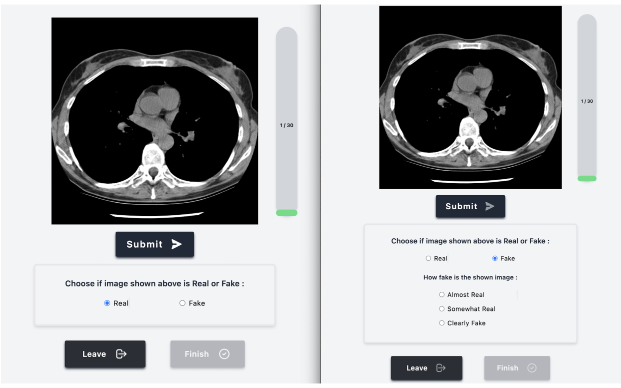
Addition of Segmentation Maps via PCA. The larger dataset contains CT scans of around 14k patients while we have segmentation maps for only 3k patients. To balance the number of segmentation maps with CT-scans for training the cycleGAN we interpolate new segmentation maps in the PCA space of existing ones. For this we used our previous methodology [2] of representing segmentation maps as a set of B-Spline curves. Since interpolations may not be perfect anatomically we use k-NN classifiers to rank the validity of segmentation maps and chose the best ones as input for training the generators in paired/unpaired training in the cycleGAN [4] setup. The creation of new segmentation maps also helps in creating CT images with novel anatomy.

3 Our Visual Turing Test for Evaluation

Some of the popular metrics generally used to evaluate generated medical images are Structural Similarity Index (SSIM), Peak Signal to Noise Ratio (PSNR), Fréchet inception distance (FID) and Inception Score (IS) among others. These metrics are a good representation of how much the generative model is able to mimic the training distribution and some metrics even give us a good idea of how much a model is able to diversify its outputs. When evaluating models that generate medical images like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Chest X-rays etc. a fundamental aspect to be considered is to verify the “medical accuracy” of the generated images. Currently, no metric can provide us with such evaluation of generative models used in medical imaging. Metrics such as FID and IS have a large dependence on the pre-trained networks which can be troublesome when the model fails to capture spatial relationships between various parts of the image. Other popular metrics such as PSNR and SSIM are numerical metrics that could be more reliable but they have been shown to be closely related to Mean Square Distance / Error (MSE) [5] for two images which is widely known to be poorly correlated with human perception of image quality or anatomical accuracy. This is a big drawback of these metrics in context of anatomical accuracy. So, we propose using the ability of humans having expertise in CT to assess our generated lung CT images and provide a better description of the generated images in the form of a Visual Turing Test for Medical Images.

Introduction - Visual Turing Test. The Visual Turing Test is a variation of the Turing Test that was first introduced by Geman et al [6] as a way to measure the level of understanding of a computer vision model. In the area of medical imaging this test was used to evaluate models based on how realistic synthetic medical images are. Chuquicusma [7] applied it to evaluate generated malignant and benign lung nodules while Han et al [8, 9] used it to evaluate brain MR images. The test is administered to human experts by showing them a randomly chosen medical image from a set of real and generated images one at a time in a random order. The expert then proceeds to give a feedback for each image shown to them without any knowledge of their actual labels. The feedback involves the experts’ opinion of whether an image is obtained from a real patient (Real) or whether it is a computer generated image (Fake). The primary idea of this test is to assess if a model is successfully able to generate medically accurate images which can be determined by measuring the number of times the model is able to fool experts into thinking that a model generated medical image is in fact a medical image obtained from a real human being. When experts are unable to separate the images into real or fake at least 50% (chance baseline) of the time, the model is said to have passed the visual Turing test.

Implementation Details. We designed a website to carry out the Visual Turing Test with a primary focus on evaluating generative frameworks that synthesize lung CT scans. The user interface for this website was created using Next.js (a server side framework built on top of react.js), tailwindCSS, Framer motion and sanity.io (GROQ Queries) as a backend to store all the responses. All responses are stored in a state which is managed by using redux, a state container. The website is hosted using vercel and is live at https://visual-turing-test.vercel.app.



Choose if image shown above is Real or Fake :
● Real ○ Fake

Choose if image shown above is Real or Fake :
○ Real ● Fake
How fake is the shown image :
○ Almost Real
○ Somewhat Real
○ Clearly Fake

Figure 4: Left: Expert user evaluation interface for our Visual Turing Test. They can choose one of "Real" and "Fake" options. Right: More options pop up if they chose "Fake".

As the test begins, the study participant is presented with an image and 2 options: "Real" or "Fake". If they choose "Fake", a sub-section pops up asking them to choose another option that best represents how fake the image looks. As shown on the right side of Figure 4 they could choose one of the "Almost Real", "Somewhat Real" and "Clearly Fake" options. After choosing a "fakeness level" the participant is shown a window as in Figure 5 where they can mark the areas that look fake in the CT image. Once they are sure of their choices the participant "submits" their feedback. We designed the test to be 30 images long so as not to overwhelm the participants. It ensures their responses are well thought out and yield an accurate measure of anatomical accuracy for our synthesized CT images. The images shown are randomly chosen from one of the three windows namely bone, lung and subdural/soft-tissue.

The test provides the following functionalities:

- Evaluates a model based on human expert feedback.
- Evaluates how close the model is to generating realistic medical images, gauging medical accuracy.
- Collects the areas of an image marked as fake by the expert, which can later be used for training better models.

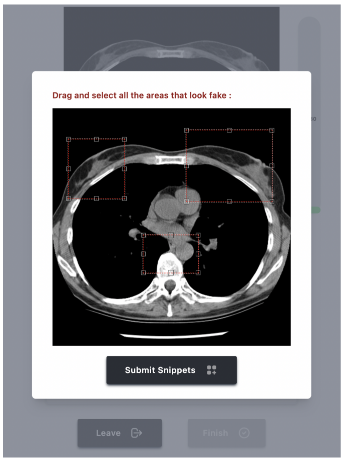

Drag and select all the areas that look fake :

Figure 5: The interface that pops up after an expert user submits the "Fake" option for an image. The red boxes indicate the areas of above image annotated by the user that he/she thinks look anatomically too inconsistent for the image to be real.

Radiologist	**Accuracy**	**TP**	**TN**	**FP**	**FN**	**Almost Real**	**Somewhat Real**	**Clearly Fake**
1-Competent	83.33%	86.67%	80%	20%	13.33%	20%	40%	20%
2-Competent	73.33%	93.33%	53.33%	46.67%	6.67%	26.67%	26.67%	0
3-Proficient	96.67%	100%	93.33%	6.67%	0	0	60%	33.33%
Average	84.44%	93.33%	75.56%	24.44%	6.67%	15.56%	42.22%	17.78%

Table 1: Statistics of responses given by 3 radiologists

Experts chosen for this test consisted of doctors, radiologists and senior radiology fellows. Since every participant will have different levels of expertise, in order to measure the performance of the model across different levels of expertise, each person is asked to indicate their medical knowledge at the end of the test.

Results. The test was taken by 3 radiologists, 1 of whom had "proficient" expertise and the others had "competent" expertise in radiology. Each of these tests presented the radiologists with 30 images belonging to 3 different windows (soft tissue, lung and bone) comprising 15 real CT images and 15 fake CT images in a random order.

Upon analyzing the responses, it was found that senior radiologists that are proficient were able to distinguish between fake and real images better than the radiology fellows who marked themselves as "competent". This can be clearly seen in confusion matrices shown in Fig. 6 where the competent test takers had difficulties in identifying all the fake images. Also, according to the table, among the chosen fake images, very few of them seemed to be "clearly fake" to the non-experts. The statistics shown in Table 1 indicate that the generative framework in consideration has not passed the Visual Turing Test as expert radiologists can easily identify most fake lung CT images from the real ones.

Analysis - The Heart Issue On closer inspection of the radiologists' feedback, we found out that expert radiologists were able to identify the fake images because of the anatomical errors in the heart. Our large lung CT dataset is entirely low-dose and has CT scans from different parts of patients' chest collected over multiple scanners. Depending on the location on a patient's chest where the CT-scan was taken, a low-dose CT image could show either 2 or all 4 chambers of the heart in a blurry fashion. Since the low-dose CT images often did not clearly show these chambers, the generator could not learn these textures. This led the synthesized images to often exhibit arbitrary number of chambers. Conversely, the other parts of the synthesized CT images including bones, muscles and the surrounding tissue looked anatomically accurate, according to the participating radiologists.

Proposed Improvements for Future Work. One of the improvements that could correct the anatomy of a heart in low-dose CT is a more focused conditional generation of the heart in a CT image. The conditional parameter could either be size, shape or a template heart image taken from a training dataset. Denoising Diffusion Probabilistic Models (DDPMs) [10] have recently shown that they could be strong candidates for high-def. conditional generation of images. Further, conditional-DDPMs like ILVR-DDPMs [11] can also be added as a refinement / extra layer for heart generation over an existing stable diffusion-based model of lung CT, giving a user more control over the synthesis process. DDPMs are more stable as compared to GANs and unlike GANs could be easily customized over existing models eliminating the need to train new models from scratch.

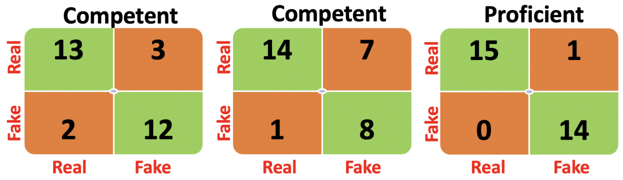


Figure 6: Confusion matrices for responses of 3 radiologists.

4 Conclusion

Our work suggests that careful implementation of texture based data augmentation combined with generative models could eliminate the "small annotated-data problem" in medical imaging domain. We also present an interactive visual turing test to evaluate these models with the help of the experts' feedback which could help develop new strategies for overcoming the shortcoming of these models.

Acknowledgements

This research was funded in part by NIH grant R01EB032716. We thank Dr. David Yankelevitz and Dr. Yeqing Zhu from the Icahn School of Medicine at Mount Sinai, NY for their help in image evaluation. We also thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

References

[1] A. Krishna and K. Mueller. "Medical (CT) image generation with style". *Fully3D* (2019).

[2] A. Krishna, K. Bartake, C. Niu, et al. "Image synthesis for data augmentation in medical ct using deep RL". *Fully3D* (2021).

[3] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *MICCAI*. 2015.

[4] S. Tripathy, J. Kannala, and E. Rahtu. "Learning image-to-image translation using paired and unpaired training samples". 2018.

[5] J.-F. Pambrun and R. Noumeir. "Limitations of the SSIM quality metric in the context of diagnostic imaging". *ICIP* (2015).

[6] D. Geman, S. Geman, N. Hallonquist, et al. "Visual turing test for computer vision systems". *PNAS* (2015).

[7] M. J. Chuquicusma, S. Hussein, J. Burt, et al. "How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis". *ISBI* (2018).

[8] C. Han, H. Hayashi, L. Rundo, et al. "GAN-based synthetic brain MR image generation". *ISBI* (2018).

[9] C. Han, L. Rundo, R. Araki, et al. "Infinite brain MR images: PGGAN-based data augmentation for tumor detection". 2020.

[10] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". *NIPS* (2020).

[11] J. Choi, S. Kim, Y. Jeong, et al. "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models". *ICCV* (2021).