Title: Quantitative Metrics for Evaluating Computed Tomography Images by Generative Adversarial Networks

Running Title: Metrics for CT-Generating GANs

*Dr. Tynan Stevens[a, b]

Finlay J. Korol[a]

[a]Nova Scotia Health, Cape Breton Regional Hospital, 1482 George St., Sydney, NS, B1P 1P3

[b]Dalhousie University, Department of Radiation Oncology, 5820 University Ave., Halifax, NS, B3H 1V7

*corresponding author

Abstract

Background

Generative adversarial network (GAN) based artificial intelligence (AI) are popular due to their capability to learn and synthesize diverse image sets. For that reason, GANs are currently being used in the domain of medical imaging, and CT image synthesis specifically sees diverse uses in cancer treatment applications. However, quantitative metrics have not been validated for medical images, making their potential utility unclear.

Purpose

This study investigates the suitability of widely used as well as recently developed GAN metrics for assessing medical image generators. A series of controlled virtual experiments were performed, along with hyperparameter tuning using quality metrics as guidance. Both observed correlation with image quality and specific failure detection were employed in assessing the utility of each metric.

Materials and Methods

GAN models using Wasserstein loss with and without gradient penalty were selected for this study due to their reputation for training stability. Training was conducted using 44416 CT images from publicly available databases. After each training run, a battery of metrics was calculated including inception score, Fréchet inception distance, sampled Fréchet distance, sliced Wasserstein distance, GAN-test/train, and likeness score. Hyperparameter tuning was performed empirically as using metrics and when necessary by inspection of generated image sets. Virtual GAN experiments were performed to assess the impact of noise, distortion, and mode collapse on each metric.

Results

Having been developed on non-medical images and lacking domain-specific adaptability, Inception Score (IS) did not perform well in the context of training a CT image generator. Distance based metrics were more sensitive to purposefully introduced image quality and diversity issues. During hyperparameter tuning, synthetic image quality and realism improved significantly.

Conclusions

Domain specific assumptions of IS render it unsuitable for medical image use. Domain agnostic metrics were more sensitive, and in particular the distance-based metrics stood out in this regard. A battery of metrics is recommended in practice for guiding the development of high-quality GANs for medical imaging studies.

Key Words: Artificial intelligence, generative adversarial network, medical imaging, computed tomography, pure image synthesis, quantitative metrics

1 Introduction

1.1 GANs

Generative Adversarial Networks (GANs) are an artificial intelligence (AI) framework developed over the last decade that have continuously gained popularity since their inception by Goodfellow et al. [1] GANs employ two neural network models in opposition engaged in a zero-sum game, typically referred to as the generator and the discriminator.[2] The goal of the discriminator is to correctly classify data— often but not limited to images — it receives as real or generated. The goal of the generator is to create data that will fool the discriminator into classifying them as real. As more training data is supplied to the GAN's discriminator, both real and generated, its ability to classify them improves, and the generator is forced to learn to make more convincing.[1]

A successful GAN eventually reaches equilibrium between generator and discriminator, at which point the generator outputs data with approximately the same distribution as training data.[1] Since the invention of GANs, there have been hundreds of varieties,[3] including the popular Wasserstein GAN (WGAN), which features a different loss function based on Earth mover's (EM) distance.[4] Additionally, instead of returning a probability of input images being real or fake, the discriminator is designed to give an unbounded score proportional to how real or fake the image is perceived to be. For this reason, the discriminator in WGANs is typically instead called a 'critic'.[4] WGANs are renowned for being more stable to train and have become important to the advancement of generative networks. A schematic diagram of a typical GAN as used in this work is shown in Figure 1.

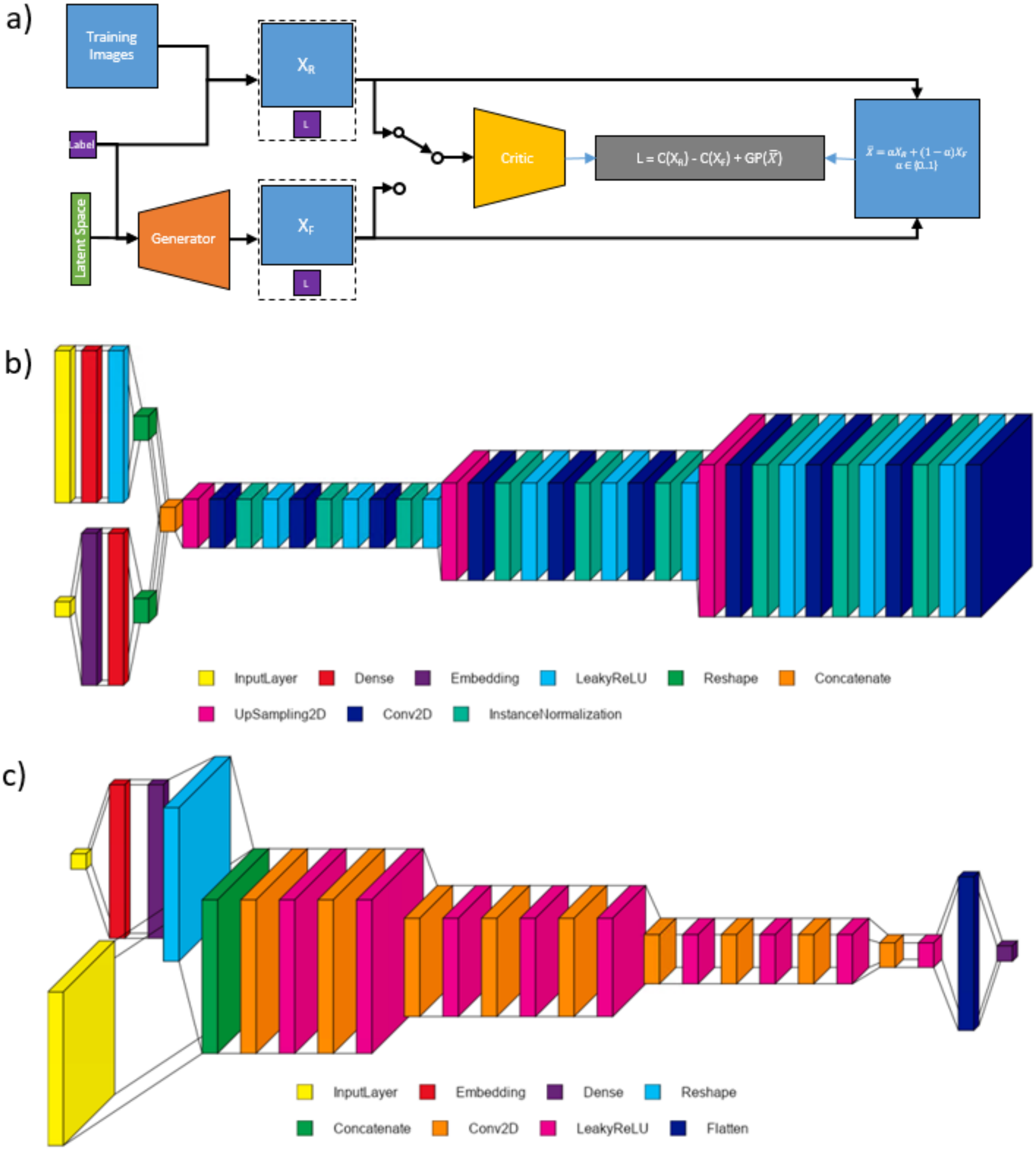

Figure 1: Network diagram of a typical cGAWN-gp implementation used in this work. In a) the overall network architecture is shown, while b) and c) show expanded views of the generator and critic respectively. The size and number of layers within the generator and critic were varied as part of the hyperparameter tuning experiments.

1.2 GANs and CT

Given the importance of Computed Tomography (CT) in medicine, there is a demand for GANs to be used in the generation of libraries of synthetic images for public to protect patient privacy.[5] Public access to a volume of these synthetic images has the potential to overcome data scarcity in areas such as education and calibration of diagnostic equipment.[6] Other CT-related applications of GANs include inter-modality image transformations between CT, and magnetic resonance (MR), ultrasound (US) or positron emission tomography (PET).[7–10] Image augmentation by means of super-resolution, artifact denoising and segmentation of images by body region have also been explored using GANs.[6] Many of these applications could be aided by an abundance of synthetic CT images, due to the inherent nature of GANs to perform better with more (high quality) training data, and the challenges of obtaining large medical imaging data sets.

Challenges in GANs

A large issue that must be addressed by any researcher working with GANs relates to their lack of associated adaptable, validated metrics.[11,12] For this reason, it is difficult to evaluate whether the generated data is realistic outside of a visual Turing test. In the context of CT image generating AI, the question addressed using a Turing test is if the generated images are indistinguishable from images taken from a real CT scanner, and would typically be performed by radiologists.[13] While Turing tests do hold merit, on their own they can struggle to detect some common problems, such as mode collapse.[14] Mode collapse occurs when a GAN produces data from a limited number of modes in the training data, making it difficult for the discriminator to identify them as generated data, despite lacking diversity.[9] Human observers would be unable to identify mode collapse reliably if asked to judge generated images individually. Furthermore, the use of expert observers is costly, and not tenable during the development and training stages of AI production.

There are two particularly common metrics used for image generating GANs that have been described extensively in the literature.[12] One is Fréchet Inception Distance (FID), which compares the mean and covariance of activations produced by real and generated data in an intermediate layer of a pre-trained image classifier network, typically the Inception v3 network. FID is intended to ensure that the real and generated images have similar distributions, and is able to detect mode collapse.[14]

The other common metric is Inception Score (IS), which uses the Inception network to classify the GAN-generated images, and calculates the Kullbach-Liebler divergence on the distribution of categories. This score aims to ensure that individual images fall into distinct categories, but also that many categories are well represented in the set.[14]

Both FID and IS use the Inception network, which is trained on ImageNet, a collection of millions of images of common objects, meaning they could give poor scores for images like CT scans regardless of potentially high GAN fidelity because the images generated do not fall into the categories used to train the Inception network.[12] FID may be more robust than IS in this regard, as it compares activations from generated images to those from real images within the same domain. In the context of CT image GANs, this ensures that the real data distribution is referenced in the metric calculation. The literature reveals that FID and IS are often acknowledged for being consistent with human perception, however this is contested.[5,15,16]

1.3 Solutions to these Challenges

As previously mentioned, a Turing test can support that a GAN is generating realistic data, however it is inadequate on its own. Another useful method to confirm that GAN-synthesized images are realistic and that their context-specific features are accurate is to compare their performance on a downstream task to that achieved on an equal quantity of real images. For example, Dikici et al. evaluated synthetic brain CT images with a brain metastasis diagnostic system, in which the false-positive rates were compared for a fixed set of data.[5]

One of GANs' strengths are their ability to synthesize data from any underlying distribution. Therefore it is important that their evaluation frameworks are not limited by unnecessary assumptions or stipulations. For GANs synthesizing medical images, metrics relying on pre-trained networks such as IS and FID must be used with caution, as poor performance can be as a result of incompatibility between the metric and the context of the GAN.[17] So-called 'domain agnostic' measures that are not pretrained, and require no labels are suitable for these purposes. These metrics extract all the information necessary to compute a score from the specific GAN's training and/or generated data.

Sampled Fréchet Distance

The sampled Fréchet distance (SFD) is a modified version of the FID, with the activations in FID being replaced by flattened vector representations of the real and generated images in SFD.[5] As a result, SFD is a potentially more flexible version of FID, and may be more applicable to images from outside the natural image domain. The SFD score includes a sum of squares difference between real and generated images, as well the trace of the real and fake image variance and covariance matrices. As with all distance-based metrics, the ideal score is 0, and no upper bound exists. Because the results from sum of squares and trace operations increase with the size of the input arrays, SFD cannot be compared directly between images generated at different resolutions.

Likeness Score

Another example of a domain agnostic metric is Likeness Score (LS).[3] LS is calculated using two measures: the intra-class distance and the between class distance, where the 'classes' in this context are the real images and the generated images. The maximum of both these distances are taken as the distance-based similarity index, which is then subtracted from 1, giving a LS bounded between zero and one.[3] LS was designed to reward creativity in the generated images (i.e. not being memorized training images), inheritance of features and styles from real images, and diversity of generated images.

GAN-train and GAN-test

Another approach used to evaluate GAN output is to train a neural network classifier on real images and apply it to the generated images to check if the categories present in the training data are readily identifiable in the generated images. Vice-versa, a classifier can be trained on the GAN output and applied to the training data. This approach is taken in the "GAN-train"/"GAN-test" framework, which can be related to the recall and precision of the GAN generator.[12] GAN-train and GAN-test are accuracy measures, and therefore are bounded between zero and one. While still flexible to domain, this approach requires the data to be labelled, and the GAN generator to be conditioned on these labels. This has two drawbacks, firstly it will not be compatible with all GAN architectures, and secondly it may require the introduction of arbitrary categorizations for medical images if obvious categories are not present (e.g. disease present vs. absent). The GAN-train and GAN-test metrics are typically analyzed in tandem, and are reported to be complementary to other metrics.[12]

Sliced Wasserstein Distance

Sliced Wasserstein Distance (SWD), like FID and SFD, is a distance-based metric, meaning its score is a distance between the real and generated images' distributions. Therefore, smaller values correspond to greater network fidelity and image quality. SWD was invented in 2011 before GANs [18], but its use in GANs has been increasing since it was used in the popular progressive growing GAN.[19] EM distance is calculated as a sum of units of work required to make two distributions equal. SWD estimates the EM distance via a sample of 1D projections of the data distribution, and aims to measure whether images in two sets have similar appearance and variation.[19]

Gap in the Literature Addressed by This Work

The current work aims to train a GAN to synthesize CT images of multiple body sections, and to measure its performance using a variety of metrics from the literature. There are a number of papers addressing generated image quality in the contexts of image denoising and reconstruction, modality transfer (i.e. between two of CT, MRI, US or PET) and image segmentation,[20] but unlike pure image synthesis, in these cases a ground truth comparison is often available.

At the time of this work, the generation of CT images using GANs was limited. While the literature has grown significantly recently, the large majority focus on CT scans targeted at single organs, especially of the lungs. Previous similar work by Park et al. has created synthetic CT images of multiple body sections, and thereafter used a Turing test conducted by expert radiologists.[21] Another study by Dikici et al. has generated CT images and has validated its GAN using SFD and diagnostic capabilities trained on real and generated images, but it is limited to the brain.[5] One study by Mann et al. generated chest CT images with GANs of COVID-19 positive and negative patients, and used a convolutional neural network (CNN) classifier on the generated images to measure the GAN's performance based on accuracy.[22] Five studies on the generation of lung nodule CT images have used a variety of validation methods, including FID, [23] a Turing test,[13,23,24] or performance in a subsequent diagnosis system.[24–26] Similarly, Frid-Adar et al. generated liver lesion CT images using a GAN and evaluated its performance using a diagnostic system.[27] To these authors' knowledge, this work will be the first in which GAN-generated CT images of multiple body regions are synthesized and quantitatively evaluated using multiple metrics.

The current study will provide groundwork to future studies in the growing field of medical image synthesizing GANs for the selection of quantitative metrics. This study will use the well-known metrics IS and FID, as well as the potential alternatives of SFD, LS, GAN-test, GAN-train, and SWD. These metrics will be validated using a series of experiments as well as network hyperparameter optimization, supporting their use in the ongoing development of GANs for CT image synthesis.

The two-network-framework used in GANs gives them a unique advantage over other AI in the field of realistic image generation.[28] Due to GANs' strengths in the generation of virtually any distribution of data, and their already widespread use in medical image synthesis, they were the clear choice in this work. It was decided that images in segments over the whole body should be used for this work for inclusiveness. The intention was to mitigate the possibility that an excluded part of the body may have seen more difficulty in metric-guided GAN training, and to ensure that these methods could be applied to whole body-CT images.

2 Materials

2.1 Training Image Acquisition

The training image set consists of 44416 images, with approximately one third each of head-neck,[29] thorax,[30] and pelvis CT images.[31] The head-neck and thorax images were acquired from the Cancer Imaging Archive,[32] which conforms to HIPAA standards. The current study follows the Cancer Imaging Archive's data usage policies and restrictions. The pelvic images from P. Liu et al were collected within IRB guidelines for sharing data.[31] As these images were all anonymized, publicly available, unlicensed and free to access and download for scientific purposes, no ethics review was required for this study.

Head-neck images all contain histologically confirmed cases of head-neck cancer.[29] Thorax images all contain cases of non-small cell lung cancer as identified using radiomics.[30] Due to the nature of the pelvic set, it is more challenging to quantify the prevalence of pathologies as it was collated using 7 smaller image sets, of which the limited details available have been collected from the associated github page. Two of these sets (178/1184 collections) contain fractures to the pelvis. 44 collections come from a set of kidney tumor CT images, 35 come from patients with either colorectal cancer or retrospective ventral hernia, 41 contain instances of cervical cancers, and two sets contained 886 various diagnoses and types of colon cancers.[31] Although these sets all contained pathologies, not all slices from each patient would show them (for example on the peripheries of a CT scan of a tumour).

Images were pre-processed by converting from DICOM to 2D PNG format, for which the Hounsfield units were converted to the integer range (0-255). A threshold of 2000 Hounsfield units was set to avoid window/level issues in the presence of metallic implants (i.e. using a fixed dynamic range of -1000 to +2000 HU). After this conversion, a small portion of images were observed to have abnormal window/level and were excluded from the training set. Images with unusual CT support structures (e.g. contrast test objects, large and bulky immobilization equipment) were also excluded. The latter exclusion criteria was justified on the basis that these features were not of interest for the desired GAN generator, and furthermore could have biased the GAN-train and GAN-test metrics by having easily identifiable features in images from one of the three categories (head-neck/thorax/pelvis). All acquired images were downsized to 64x64 to improve GAN training speed.

3 Methods

3.1 Network Design

The network employed for the current study is a conditional WGAN with gradient penalty (cWGAN-GP). While the number and size of layers varied in the process of hyperparameter tuning, a representative network architecture is shown in Fig. 1. In general, the generators consisted of two input layers for the latent space vector and conditional label (head-neck/thorax/pelvis), both of which were densely connected and reshaped into a layer matching the minimum convolution layer size (e.g. 8x8). These two layers were concatenated and then fed into successive layers of upsampling and convolution/instance normalization/leaky ReLU activation, until the desired image size (64x64) was reached. Likewise the critic takes two input layers (an image and conditional label), with the label being fully connected to a layer of the same shape as the input image, concatenated, and fed to successive downsampling, convolution and leaky ReLU layers. Once the minimum layer size was reached, the final layer was flattened and densely connected to a linear output layer. No instance normalization was used in the critic. The training loss included three terms as in Gulrajani et al.'s gradient penalty implementation:[33] Wasserstein loss for the critic on real images, Wasserstein loss for the critic on fake images, and gradient penalty loss on weighted averages of real and fake images. 'Vanilla' conditional GAN (cGAN) and conditional GAN with Wasserstein loss (cWGAN) were also tested. While all three architectures were capable of producing high quality generators, the authors noticed a substantial improvement in stability for larger numbers of training epochs with cWGAN-GP.

All networks were developed in Python, utilizing the Keras (v2.4.3) and TensorFlow (v2.4.1) packages. Training was conducted on a workstation equipped with an AMD threadripper 3970X CPU, 128 GB of system RAM, and four NVIDIA RTX 3090 GPUs with 24GB of VRAM each.

3.2 Virtual GAN Experiments

Before applying the GAN quality metrics in the context of hyperparameter tuning, a series of experiments on controlled GAN failure modes and image quality deteriorations was performed. None of these experiments used generated images. In each experiment, a sample of 2000 real (representing GAN-training) images from the GAN training set and 2000 manipulated (representing GAN-generated) images were produced. These real and manipulated samples were then used to compute each of the seven GAN metrics described previously (IS, FID, SFD, LS, GAN-test, GAN-train, and SWD). Note that IS does not require the real image sample, and the GAN-test/train additionally requires the class labels. All other metrics require only the real and manipulated image samples. Each experiment was performed for three repetitions.

Two types of simulated mode collapse experiments were performed: partial mode collapse (PMC) and multimode collapse (MMC). PMC featured one image from each category (head-neck, thorax, pelvis-upper thighs), not included in the real image sample, repeated to create the mode collapsed portion. The proportion of mode collapsed images was incremented from 1/16th of the manipulated image set up to 'total mode collapse', which consisted of just the three images (one per category) replicated until the set was 2000 in length. For the MMC experiments, total mode collapse was the same as above, but the number of modes represented was gradually increased up to 16 samples from each category (i.e. approximately 40 copies of 48 different images in total). Again, none of the images in the real image sample were used in the MMC set.

Partial memorization (PMe) of the real image set was also investigated in a virtual experiment. For this test, a growing fraction of the real image sample was duplicated in the 'manipulated' set, with the remainder being unique real image samples. Total memorization was modelled by fully duplicating the real image set as the manipulated set. PMe started with 1/16th memorized images, doubling the number of memorized images until total memorization.

The next set of experiments modelled incremental distortions to image quality. Distortions were added to 2000 images using Python to create the manipulated sets, which were again unique from the real image sample. The distortions used were a swirl effect, gaussian noise, salt and pepper noise, gaussian blur and implanted black squares as in Heusel et al.'s supporting experiments to FID,[34] as well as a square swap effect (see Figure 2). In each of these experiments, five increasing distortion strengths were applied to the manipulated images. A baseline optimal value for each metric was acquired using un-distorted images for the manipulated set.

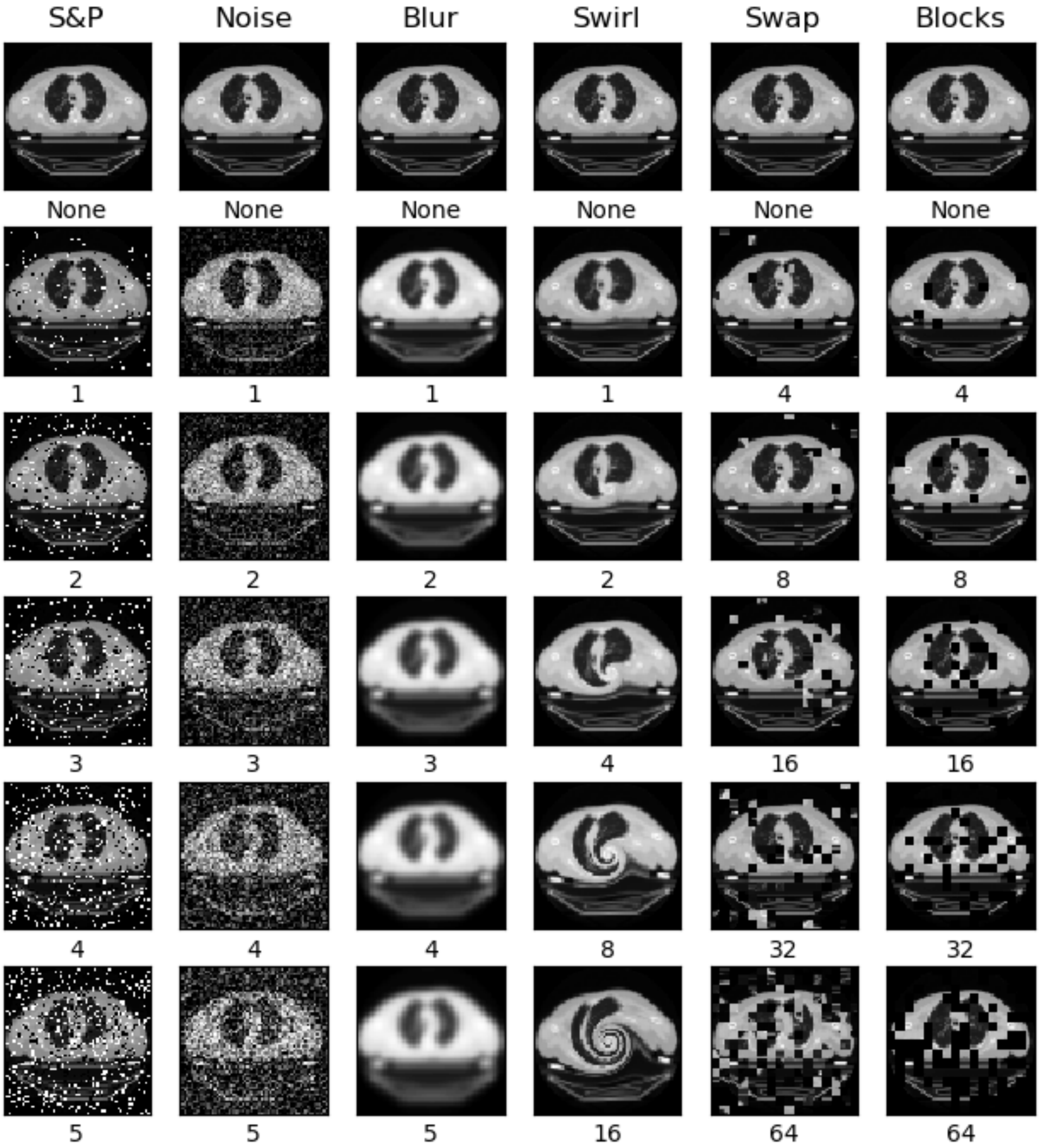

Figure 2: Example images generated for the virtual GAN experiments. The first two columns demonstrate the introduction of varying levels of either salt and pepper (S&P) or Gaussian type noise. The Third column demonstrates repeated application of a 5x5 Gaussian smoothing kernel. The fourth column shows the effects of a spiral or swirl distortion, and the final two columns show increasing numbers of swapped or blacked out square patches. The number under each image indicates the number of times the disturbance was applied, or in the case of the swirl artifact the severity parameter.

3.3 Hyperparameter Tuning

We performed hyperparameter tuning to assess the ability of each metric to guide the optimization process. The goal of this process was not to arrive at the best possible network configuration, but to evaluate the use of the metrics in a real-world setting, responding to realistic training challenges such as partial mode collapse and marginal improvements in image quality between configurations.

The hyperparameters to be tuned in this network were size of the latent space, size of the label embedding space, size of the generator and critic convolution layers, size of the generator and critic convolution kernels, minimum image size for upsampling/downsampling, number of convolutions per upsampling/downsampling network layer, number of fully connected layers in the critic, the number of learning iterations for the generator and critic, the number of images per training batch, the alpha value for the network's leaky ReLU activation layer, the learning rates of the generator and critic, and the gradient penalty weight.

Hyperparameter tuning was guided by aiming to optimize the values of all metrics. For each hyperparameter a list of potential values were chosen, and the GAN was trained for 10 epochs. The range of values tested was chosen to include popular settings from the literature, and where possible edge cases that would be expected to push the limits of the GAN model. Visual inspection was used, especially in cases where the metrics did not clearly favour one hyperparameter value. When two values of a hyperparameter were indicated as best by different metrics, those GANs were re-trained for 20 epochs, followed by another round of metrics comparisons and visual inspection until a decision could be reached (i.e. favoring training stability).

This coarse preliminary tuning was performed on all hyperparameters, followed by a round of fine tuning. For fine tuning, values were chosen on either side of the optimal setting from the coarse stage and the GAN model was re-trained for these intermediate values following the same strategy as above.

4 Results

4.1 Virtual GAN Experiments

We found during GAN hyperparameter tuning that inception score was insensitive to even quite dramatic changes in visually perceived quality. Moreover, we observed that the IS for real CT images (not used during GAN training) was not substantially better, indicating that the omission of ‘real image’ data and reliance on the inception network object categories renders IS unsuitable for medical image GANs. This was reinforced in the virtual GAN experiments, where we observed IS to be relatively insensitive to even severe image distortion (Figure 3). IS was even observed to increase with the introduction of Gaussian noise or low levels of blacked out squares. This metric was thus excluded from further analysis, and the authors discourage it’s use for medical image GANs.

The partial memorization virtual experiment showed (Figure 4) that the distance based metrics all converge on 0 when the manipulated set is a copy of the real set (ie. fully memorized). Likewise the fully unmemorized set represents an ideal score, where the manipulated images are unaltered CT images, unique from the training set. This provides an estimate of the lower bound for distance based metrics, below which memorization may be indicated. Likeness score similarly achieves a ‘perfect score’ of 1.0 for the fully memorized case, whereas the GAN-test/train metrics don’t achieve perfect classification even when the real and manipulated data are identical, owing to the stochastic nature of training a CNN classifier.

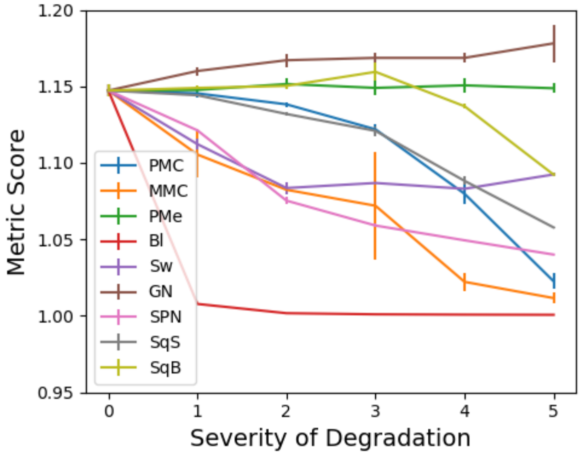

Figure 3: Inception Score (IS) over varying degrees of introduced distrortions. PMC = Partial mode collapse, MMC = Multi-mode collapse, Bl = gaussian blur, Sw = swirl artifact, GN = Gaussian noise, SPN = Salt and Pepper noise, SqS = swapped squares, SqB = Blacked Squares.

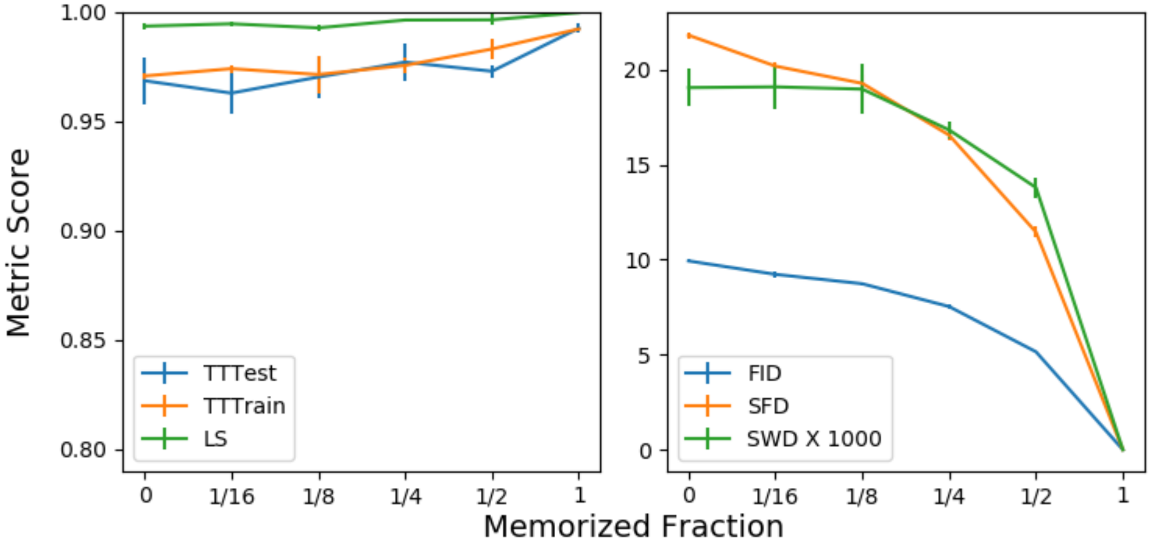

Figure 4: GAN metrics as a function of training set memorization fraction. Distance based metrics converge to a perfect score of 0 for full memorization, whereas LS, and the test/train scores converge towards a perfect score of 1. LS = likeness score, TT = test/train, FID = Fréchet inception distance, SFD = sampled Fréchet distance, SWD = sliced Wasserstein distance.

The mode collapse experiments revealed that all metrics except the 'test' portion of GAN-test/train respond to mode collapse (Figure 5). However, LS and GAN-test/train appear less sensitive to partial mode collapse, not deviating significantly from the score given for the fully uncollapsed real set until half or more of the images in the manipulated set are identical to one another. By comparison the distance based metrics began to respond to as little as a 1/8th mode collaposed set. All metrics but GAN-test responded strongly to the fully memorized set. This can be understood by noting that this score uses a classifier trained on the (diverse) training set, which needs to only correctly classify a small number of mode collapsed test images. Conversely, the 'Train' score suffers because the classifier trained on a mode collapsed image set fails to capture the diversity present in the training set. Similar results were obtained for the PMC and MMC experiments, though MMC was more impactful on the metrics, as image diversity decreases more quickly with the mode collapsed fraction.

A similar trend is observed for the degradation and distortion experiments, with the distance-based metrics responding to even small perturbations of any kind (Figure 6). GAN-test/train were relatively insensitive to Gaussian blur, whereas FID was comparatively strongly impacted. SFD was on average the most sensitive of the distance metrics to the geometric distortions (swirl, masked or swapped patches). LS also responded to all artifacts, although less so to the swirl distortion than others, despite the visually striking nature of this effect. Overall, all metrics demonstrated the ability to detect these incremental changes in image quality.

The final virtual experiment focussed on the introduction of noise to the test image set (Figure 7). Increasing levels of gaussian noise had a dramatic effect on the distance based metrics, and FID in particular. By comparison GAN-test/train and LS were again insensitivte to Gaussian additive noise, although the latter did show a clear trend of decreasing LS with increasing noise content. Salt and pepper noise was more impactful on all metrics except GAN-train, and had the greatest effect of all artifacts tested on the LS and SFD metrics.

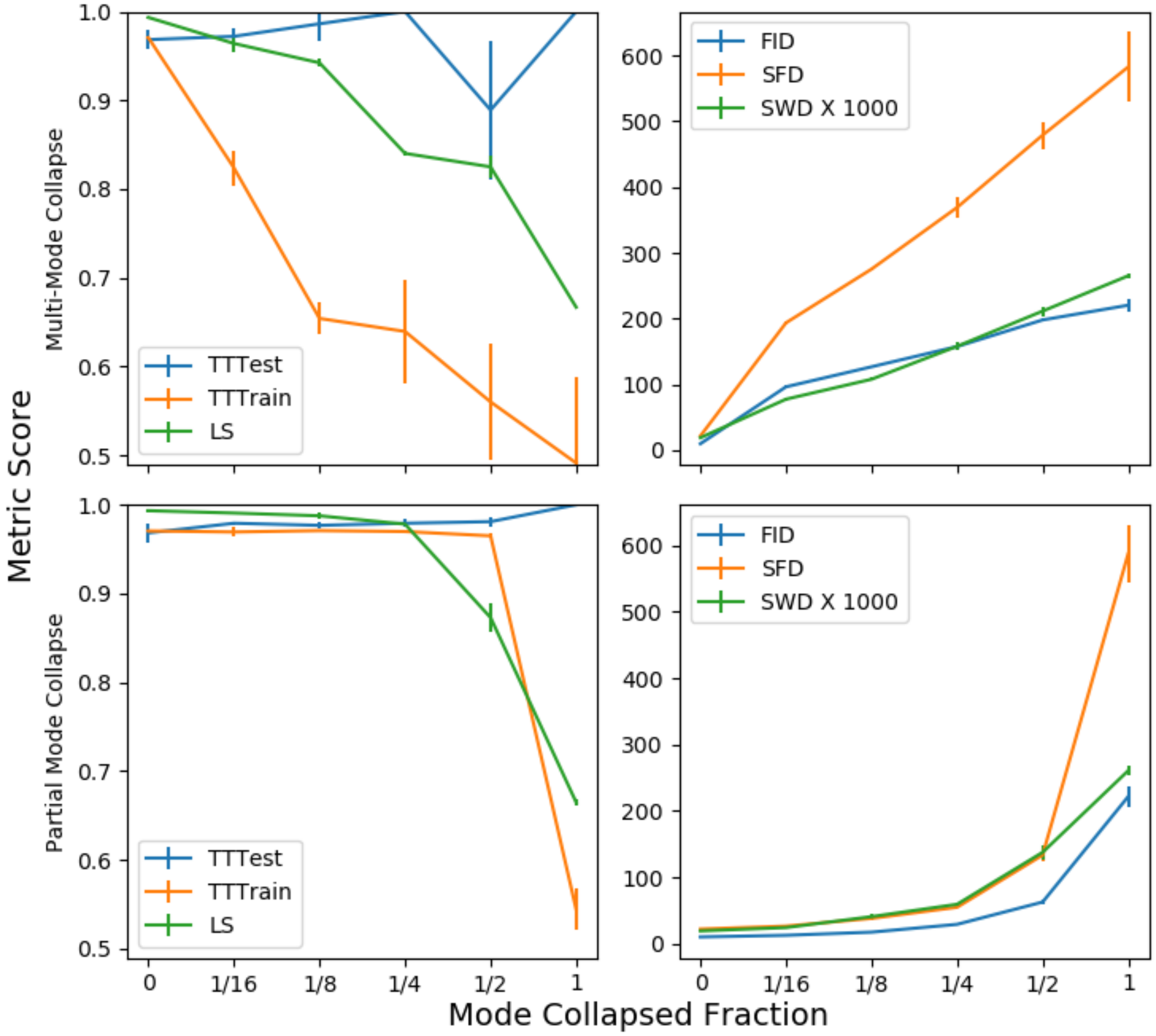

Figure 5: Summary of virtual GAN experiments simulating mode collapses.

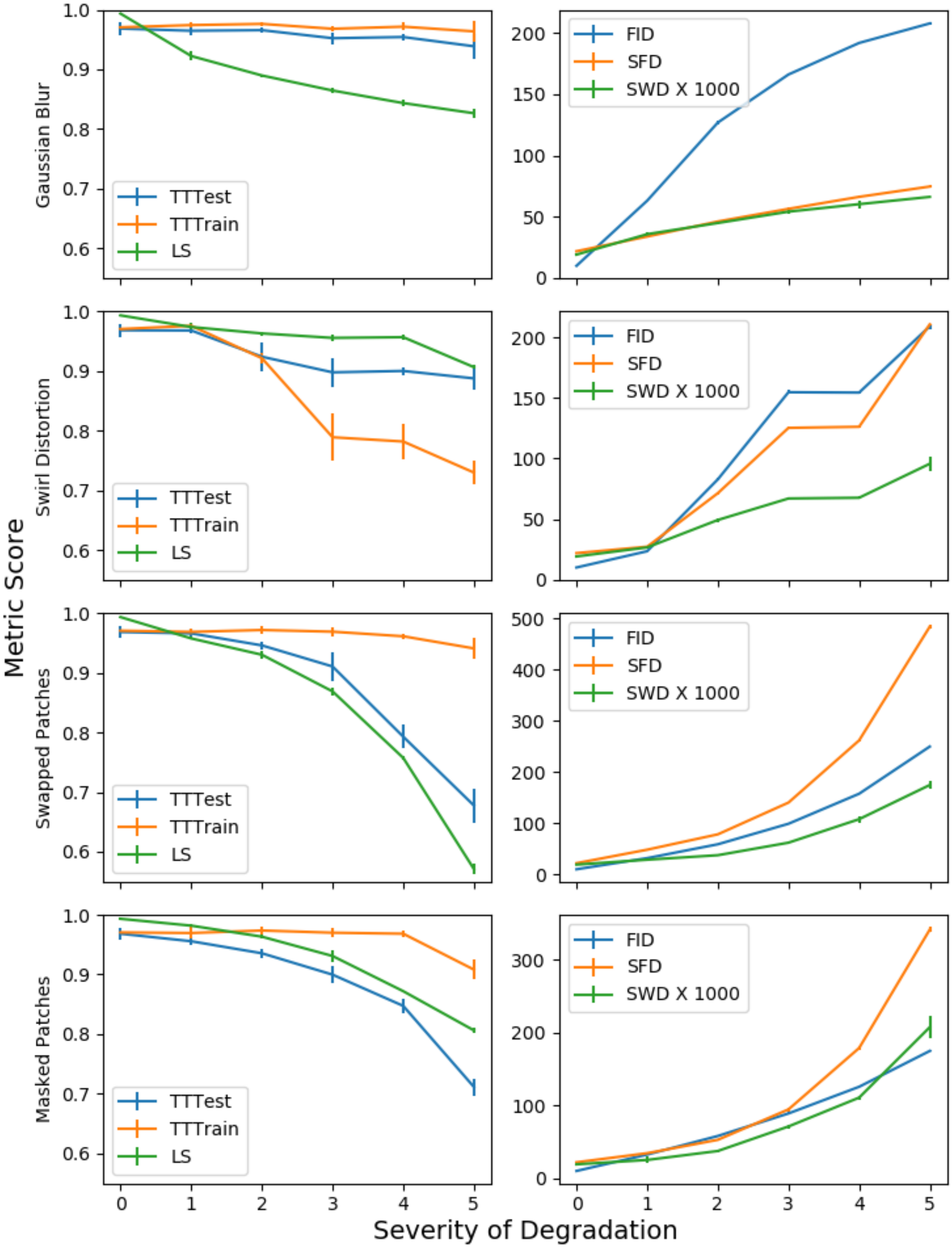

Figure 6: Summary of virtual GAN experiments simulating image distortions.

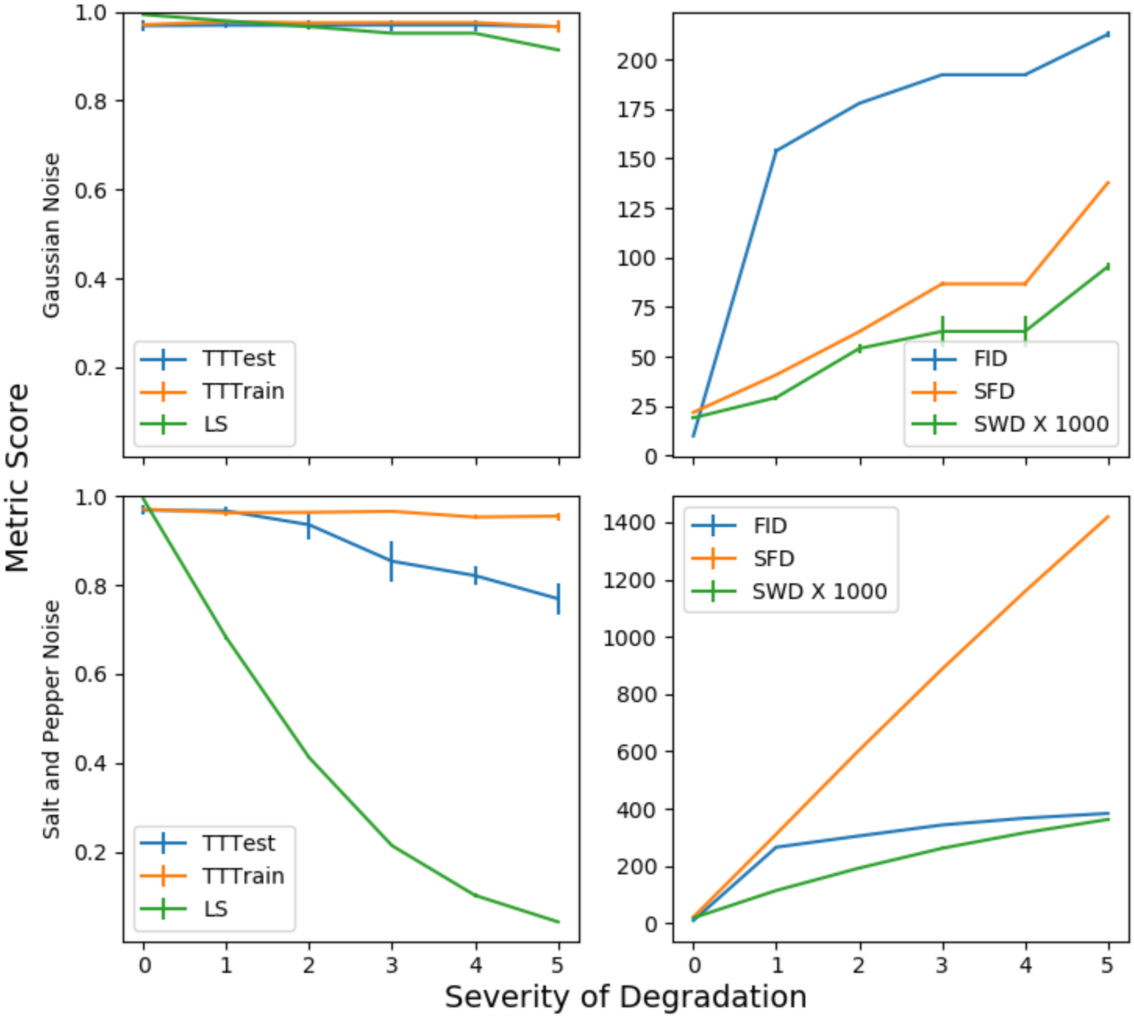

Figure 7: Summary of virtual GAN experiments simulating increasing image noise.

4.2 Hyperparameter Tuning

While there was general agreement between metrics at the extreme ranges of hyperparameters, tuning frequently revealed subtle differences between the various metrics when used in practice. For example, Figure 8 shows the effect of tuning the minimum convolution layer size. Using a size of 64 effectively eliminates all convolutional layers, reducing the network by fully connecting it, causing all metrics to suffer. On the other end of the spectrum, some metrcs indicated loss of GAN quality for a minimum layer size of 4 (corresponding to 4 levels of downsampling and convolution), which may simply reflect insufficient training epochs to achieve quality output with the additional network parameters introduced. With gradient penalty enabled, the metrics improved nearly uniformly for minimum layer sizes of 4 to 32 pixels (corresponding to 1 to 4 downsampling steps). This latter observation was consistently noticed in tuning of other hyperparameters as well, with the gradient penalty term improving the robustness of the GAN training to poor hyperparameter selection.

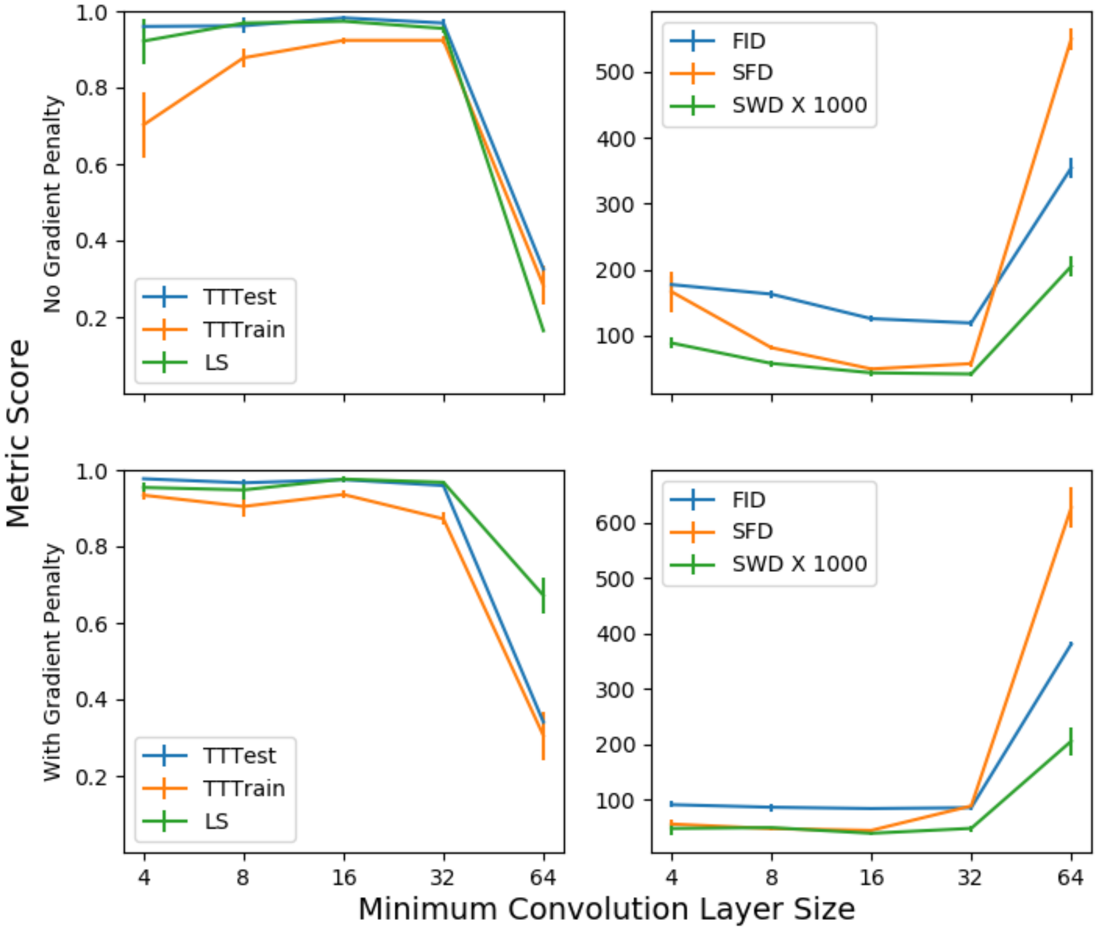

Figure 8: GAN quality metrics over tested values of the minimum convolution layer size with/without gradient penalty in the loss function.

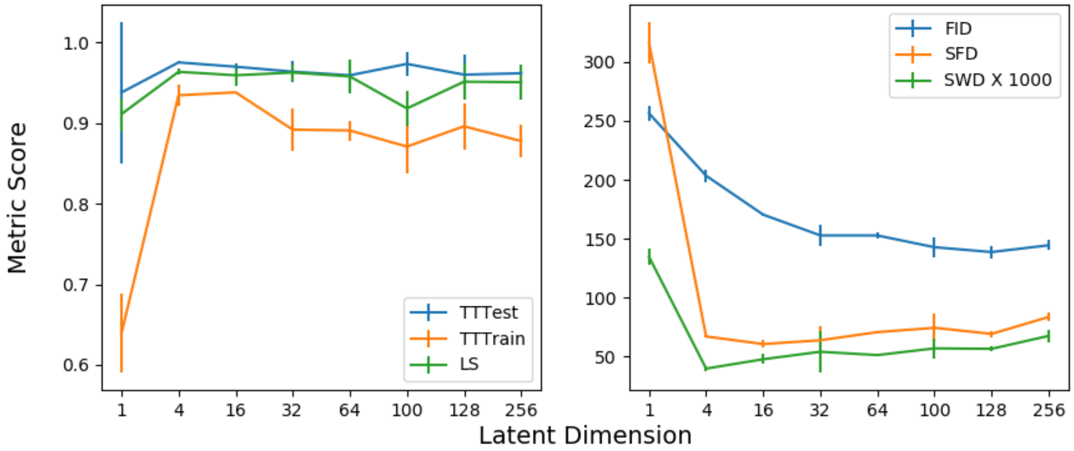

Figure 9: GAN quality metrics as a function of latent dimension size.

While for most of the hyperparameter optimization tests performed there was a similar trend between the metrics, the optimal settings frequently differed based on the metric chosen. The best example of this was the optimal latent space size (LSS), which was different for SWD (LSS=4), SFD (LSS=8), GAN-train (LSS=16), LS (LSS=32), GAN-test (LSS=100) and FID (LSS=128). However as shown in Figure 9, for several metrics there were a number of values of this hyperparameter that were within experimental erorr of one another, and the effects on the generated images were subtle when assessed qualitatively. Notably there was often clear concensus on poor parameter choices, as the use of a very small latent space dimension universally produced the worst metric scores. The metrics tested also generally agreed on a diminishing returns for increasing latent dimension, with the distance based metrics even demonstrating a trend to degrade for very large latent space sizes.

5 Discussion

In this work, we have demonstrated the application of several popular GAN quality metrics in the application to CT generation for medical applications. Overall, we have shown that these metrics are sensitive to purposefully introduced artifacts and noise, as well as to various types of memorization or mode collapse. The exception to this trend was IS, which was relatively insensitive, and even showed misleading results. This is likely because IS rewards generators that produce high quality images spanning multiple categories of natural objects, with no reference to the real images from a given application-specific domain like CT. That IS was seen to increase with some purposefully introduced artifacts may reflect that these artifacts created images with increased diversity from the viewpoint of the Inception v3 model. For these reasons, the IS should not be relied on to guide parameter or network design choices for medical images.

Most of the other metrics tested were found to correlate with perceived image quality in the majority of the virtual and hyperparameter tuning experiments. However, there were substantial differences between the various metrics on any given test, often with no clear consensus on the best hyperparameters at the fine-tuning scale. As each metric responds to different aspects of image quality and diversity, it is not possible to universally recommend one over the others, and in practice using a battery of metrics is likely to be more appropriate for the wide range of real world challenges faced when training GANs. For example, SFD was highly sensitive to mode collapse, but for images lacking sharp details (the Gaussian blur experiment), likeness score and FID were more sensitive metrics. Further developing such a battery of metrics, with each tailored to detect particular GAN deficiencies, could help to more directly guide GAN development.

The use of a separate classifier neural network to assess the quality of GAN output relative to the training data has some unique challenges, as shown with the GAN-test/train metrics used in this work. For one, the data must have categorical labels available, which is not the case for all applications. Secondly, the metric scores will be dependent on all features of the classifier including network design, architecture, training parameters, and input image shapes, rendering comparisons between implementations challenging or impossible. Finally, these metrics were the least sensitive (excluding IS) in many of our tests. This speaks to the robustness of neural network classifiers to image noise and degradation, but unfortunately renders them difficult to use in guiding nuanced choices of network design and parameters.

An important observation made during the hyperparameter optimization experiments is that the metrics only provide a snapshot in time of the output quality. There is no guarantee that the network paramaters that produce the highest quality images after 10 epochs will continue to perform best after 100 or 1000 epochs, as GAN training is well known to be prone to collapse. While outside the scope of this study, we noted that the cWGAN-GP architecture was more robust in this regard than cGAN or cWGAN, and that early success indicated by the tested metrics more often predicted later success as well. A future study should investigate the degree to which the various metrics demonstrate convergence over training epochs for long training runs, as this may be particularly helpful in comparing CNN architectures.

While no other study has examined the suitability of multiple GAN quality metrics for medical images, there are a few key pieces of literature against which we can compare our results. Dikici et al. defined the SFD metric in their work generating brain metastatis MR images. They used this metric to validate individual GANs selected to include in an ensemble of generators, which was then subsequently validated using an automated diagnosis algorithm, demonstrating similar performance on real and fake data. While not a direct validation of SFD, there is an implicit assumption that the metric is valid since the selected generators produce meaninfgul data for downstream uses. We also found SFD to perform quite well across our virtual experiments.

Finally, it was noted that some metrics are more computationally expensive than others – for instance the need in SFD to compute variance/covariance matrices and their inverses is quite onerous for large images. In practice, down-sampling to an intermediate resolution prior to SFD calculation is thus necessary. Similar is implicitly done in the calculation of FID, as the images must be rescaled to match the inception network input image size.

6 Conclusions

This work demonstrated the utility of several GAN quality metrics for guiding network design and training. The LS, FID, SFD and SWD metrics had attractive combinations of responsiveness and ease of implementation. The distance-based metrics have the advantage of extended dynamic range, which renders them less prone to saturation as image quality of the generated images approach that of the training set, but care must be taken to avoid rewarding memorization. As no single metric was clearly better than others, a battery of tests approach is recommended in practice.

7 Conflicts of Interest

The authors have no relevant conflicts of interest to disclose.

6 References

1. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. *ArXiv14062661 Cs Stat*. Published online June 10, 2014. Accessed July 8, 2021. http://arxiv.org/abs/1406.2661

2. Wang T, Lei Y, Fu Y, et al. A review on medical imaging synthesis using deep learning and its clinical applications. *J Appl Clin Med Phys*. 2021;22(1):11-36. doi:10.1002/acm2.13121

3. Guan S, Loew M. A novel measure to evaluate generative adversarial networks based on direct analysis of generated images. *Neural Comput Appl*. 2021;(2021). doi:https://doi.org/10.1007/s00521-021-06031-5

4. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. In: *International Conference on Machine Learning*. PMLR; 2017:214-223. Accessed August 31, 2021. https://proceedings.mlr.press/v70/arjovsky17a.html

5. Dikici E, Bigelow M, White RD, Erdal BS, Prevedello LM. Constrained generative adversarial network ensembles for sharable synthetic medical images. *J Med Imaging*. 2021;8(2):024004. doi:10.1117/1.JMI.8.2.024004

6. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. *Med Image Anal*. 2019;58:101552. doi:10.1016/j.media.2019.101552

7. Alvarez Andres E, Fidon L, Vakalopoulou M, et al. Dosimetry-Driven Quality Measure of Brain Pseudo Computed Tomography Generated From Deep Learning for MRI-Only Radiation Therapy Treatment Planning. *Int J Radiat Oncol*. 2020;108(3):813-823. doi:10.1016/j.ijrobp.2020.05.006

8. Kazemifar S, McGuire S, Timmerman R, et al. MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. *Radiother Oncol*. 2019;136:56-63. doi:10.1016/j.radonc.2019.03.026

9. Kazemifar S, Montero AMB, Souris K, et al. Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors. *J Appl Clin Med Phys*. 2020;21(5):76-86. doi:10.1002/acm2.12856

10. Tang B, Wu F, Fu Y, et al. Dosimetric evaluation of synthetic CT image generated using a neural network for MR-only brain radiotherapy. *J Appl Clin Med Phys*. 2021;22(3):55-62. doi:10.1002/acm2.13176

11. Borji A. Pros and cons of GAN evaluation measures. *Comput Vis Image Underst*. 2019;179:41-65. doi:10.1016/j.cviu.2018.10.009

12. Shmelkov K, Schmid C, Alahari K. How Good Is My GAN? In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018*. Lecture Notes in Computer Science. Springer International Publishing; 2018:218-234. doi:10.1007/978-3-030-01216-8_14

13. Chuquicusma MJM, Hussein S, Burt J, Bagci U. How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis. *ArXiv171009762 Cs Q-Bio*. Published online January 8, 2018. Accessed July 8, 2021. http://arxiv.org/abs/1710.09762

14. Xu Q, Huang G, Yuan Y, et al. An empirical study on evaluation metrics of generative adversarial networks. *ArXiv180607755 Cs Stat*. Published online August 16, 2018. Accessed July 8, 2021. http://arxiv.org/abs/1806.07755

15. Liu S, Wei Y, Lu J, Zhou J. An Improved Evaluation Framework for Generative Adversarial Networks. *ArXiv180307474 Cs*. Published online July 19, 2018. Accessed July 8, 2021. http://arxiv.org/abs/1803.07474

16. Zhou S, Gordon ML, Krishna R, Narcomey A, Fei-Fei L, Bernstein MS. HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models. *ArXiv190401121 Cs*. Published online October 31, 2019. Accessed July 13, 2021. http://arxiv.org/abs/1904.01121

17. Grnarova P, Levy KY, Lucchi A, et al. A domain agnostic measure for monitoring and evaluating GANs. *ArXiv181105512 Cs Stat*. Published online July 15, 2020. Accessed July 13, 2021. http://arxiv.org/abs/1811.05512

18. Julien R, Peyré G, Delon J, Marc B. Wasserstein Barycenter and its Application to Texture Mixing. In: *SSVM'11*. Springer; 2011:435-446. Accessed August 30, 2021. https://hal.archives-ouvertes.fr/hal-00476064

19. Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *ArXiv171010196 Cs Stat*. Published online February 26, 2018. Accessed August 30, 2021. http://arxiv.org/abs/1710.10196

20. Sorin V, Barash Y, Konen E, Klang E. Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review. *Acad Radiol*. 2020;27(8):1175-1185. doi:10.1016/j.acra.2019.12.024

21. Park HY, Bae HJ, Hong GS, et al. Realistic High-Resolution Body Computed Tomography Image Synthesis by Using Progressive Growing Generative Adversarial Network: Visual Turing Test. *JMIR Med Inform*. 2021;9(3):e23328. doi:10.2196/23328

22. Mann P, Jain S, Mittal S, Bhat A. Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks. *ArXiv210511241 Cs Eess*. Published online May 20, 2021. Accessed July 12, 2021. http://arxiv.org/abs/2105.11241

23. Semiletov A, Vatian A, Krychkov M, et al. Comparative Evaluation of Lung Cancer CT Image Synthesis with Generative Adversarial Networks. In: Paszynski M, Kranzlmüller D, Krzhizhanovskaya VV, Dongarra JJ, Sloot PMA, eds. *Computational Science – ICCS 2021*. Lecture Notes in Computer Science. Springer International Publishing; 2021:593-608. doi:10.1007/978-3-030-77967-2_49

24. Nishio M, Muramatsu C, Noguchi S, et al. Attribute-guided image generation of three-dimensional computed tomography images of lung nodules using a generative adversarial network. *Comput Biol Med*. 2020;126:104032. doi:10.1016/j.compbiomed.2020.104032

25. Onishi Y, Teramoto A, Tsujimoto M, et al. Automated Pulmonary Nodule Classification in Computed Tomography Images Using a Deep Convolutional Neural Network Trained by Generative Adversarial Networks. *BioMed Res Int*. 2019;2019:e6051939. doi:10.1155/2019/6051939

26. Toda R, Teramoto A, Tsujimoto M, et al. Synthetic CT image generation of shape-controlled lung cancer using semi-conditional InfoGAN and its applicability for type classification. *Int J Comput Assist Radiol Surg*. 2021;16(2):241-251. doi:10.1007/s11548-021-02308-1

27. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. 2018;321:321-331. doi:10.1016/j.neucom.2018.09.013

28. Lan L, You L, Zhang Z, et al. Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Front Public Health*. 2020;8. Accessed January 18, 2022. https://www.frontiersin.org/article/10.3389/fpubh.2020.00164

29. Vallières M, Kay-Rivest E, Perrin L, et al. Data from Head-Neck-PET-CT. Published online 2017. doi:10.7937/K9/TCIA.2017.8OJE5Q00

30. Aerts HJWL, Wee L, Rios Velazquez E, et al. Data From NSCLC-Radiomics. Published online 2019. doi:10.7937/K9/TCIA.2015.PF0M9REI

31. Liu P, Han H, Du Y, et al. Deep Learning to Segment Pelvic Bones: Large-scale CT Datasets and Baseline Models. *ArXiv201208721 Cs*. Published online March 31, 2021. Accessed August 31, 2021. http://arxiv.org/abs/2012.08721

32. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:https://doi.org/10.1007/s10278-013-9622-7

33. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved Training of Wasserstein GANs. *ArXiv170400028 Cs Stat*. Published online December 25, 2017. Accessed September 2, 2021. http://arxiv.org/abs/1704.00028

34. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *ArXiv170608500 Cs Stat*. Published online January 12, 2018. Accessed October 14, 2021. http://arxiv.org/abs/1706.08500