Towards a Conditional Generative Model for Trajectory Modeling in Medical Imaging

ECSE 556 Final Project Report

Anita Kriz
260823730

Abstract

There is a need for the controllable generation of medical images in order to monitor chronic diseases such as Multiple Sclerosis (MS). In order to develop such models, highly correlated attributes need to be sufficiently separable so that the model can learn how each independently effects the generated image. Learning a disentangled representation of attributes would thus allow for specific control in the generation process. In this work, the latent spaces of the StyleGAN are studied to understand how disentangled they are. Disentanglement is both qualitatively and quantitatively assessed for the StyleGAN and its variations, with metrics formulated from models specifically trained on MS data. Although the resulting generation still requires optimization, this is a first step in creating a conditional model for controllable generation in medical imaging with vast applications such as trajectory modeling and treatment effect.

1 Introduction

Medical imaging is an important tool for diagnosing and monitoring progression in chronic diseases such as Multiple Sclerosis (MS) (1). Although extensive training and practice is required of radiologists to read images such as MRIs, studies show that they can and sometimes will get it wrong, with approximately 40 million diagnostic errors involving imaging occurring annually worldwide (2) (3). With such a high prevalence of errors in diagnosis, research directions such as trajectory modeling of disease and personalized medicine become infeasible without alternative methods. Due to the growing demands of radiologists and the challenges in interpreting medical images, deep learning is a field that is actively being explored to develop an automatic, robust, and standardized methods for problems such as these.

Generative AI - a type of artificial intelligence (AI) that is able to create new and original content - has had several successes in the past few years. Powerful examples include Chat-GPT or DALL-E, both which use user prompts to return original text and image generations, respectively. Given these rapid advancements, the utilization of generative AI in the field of medical imaging holds significant potential. However, in the medical field, random and uncontrolled generation via these black-box models is largely undesired due to concerns such as trustworthiness and fairness. Moreover, important questions pertaining to specific patients or treatments cannot be answered in this way. Thus, the gap in the field emerges in the question: how can we harness control over generative models while being able to generate clinically relevant images?

This concept is known as *conditional* generation. More specifically, the ability to change certain attributes (e.g. extent of disease, treatment, patient identity) while keeping the other details unchanged is a requirement(4)(5). An illustrative example would be a doctor who wants to use such a model to know how a patient would look like if they were healthy. Via conditional generation, they should be able to fix the patient's identity (and any other non-relevant attribute) while toggling the disease extent. Such generation would allow for understanding the changes induced to the brain by the disease (6). In all, conditional generation has a vast number of applications in the medical imaging field, including understanding disease progression, monitoring treatment effect, and unlocking personalized medicine.

Although this is a newly possible and exciting concept in modern medicine, developing such models in medical imaging does not come without its constraints. Given that datasets used in medical imaging are often from clinical trials, there is a high correlation between certain attributes. For example, as a patient is in a clinical trial for longer (i.e as they age), their disease tends to worsen. However, this does not mean that aging causes disease progression, or vice versa. With the bias present in the dataset, it may be difficult for a model to fix one attribute while altering another without causing modifications to the image that are related to the fixed attribute. Therefor, there is a need to "disentangle" these highly correlated attributes from one another, such that the a generative model is able to learn what changes should correspond to which attribute. Once sufficiently disentangled, it should be possible to condition on attributes effectively. Unfortunately, disentanglement does not come without its limitations: as found in recent work, disentangling representations of images often times lead to reduced generation quality (7).

In this work, we use a form of Generative Adverserial Networks (GANs), the StyleGAN, to generate brain images (8). Unlike GANs, StyleGANs use an intermediate latent space before the generation process in order to allow for a more disentangled representation. To understand how disentangled the attributes are in this generator architecture, we use two metrics as proposed by the authors of StyleGAN, perceptual path length (PPL) and linear separability. We then compare the metrics using this architecture to traditional GANs. The contributions of this work are:

1. The formulation of disentanglement metrics by training models specific to MS training data
2. A qualitative and quantitative comparison of the disentanglement of latent spaces in StyleGAN and GAN architectures

With these contributions, we come one step closer to a conditional generative model in the field of medical imaging.

2 Background

2.1 Generative Adversarial Network (GAN)

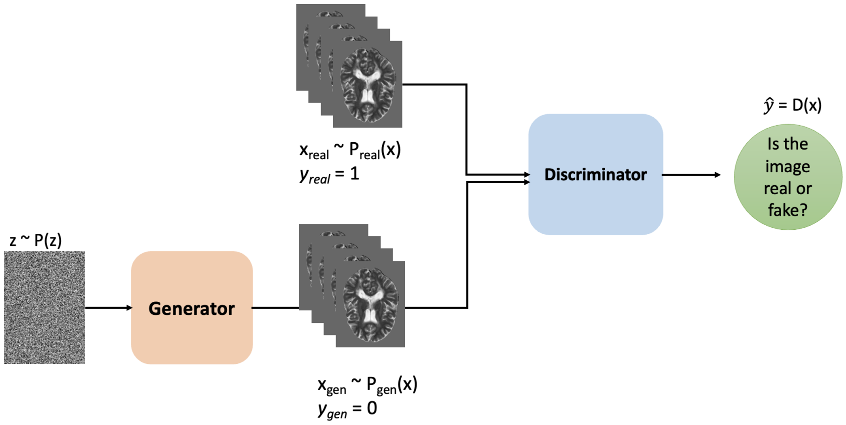
A generative adversarial network (GAN) (9), is a type of generative model that relies on two key components: a generator and a discriminator. As shown in Figure 1, the generator takes a randomly sampled vector z from a set distribution (often a Gaussian) of noise and generates images $x_{\text{gen}} \sim P_{\text{gen}}$. Throughout training, it attempts to generate images such that $P_{\text{gen}}(x)$ is as close to $P_{\text{real}}(x)$ as possible. The discriminator then takes either the real or generated image and decides whether that image is coming from the real or fake distribution. Learning is done via the MiniMax loss function as shown here, where the discriminator tries to maximize it and the generator tries to minimize it:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{x \sim P_{\text{data}}}[\log(1 - D(G(x;\theta);\phi))] + \mathbb{E}_{z \sim P_z}[\log(D(G(z;\theta);\phi))] \tag{1}$$

Thus throughout training, the discriminator gets better at distinguishing between real and fake data, and the generator learns to fool the discriminator. Eventually, the discriminator can be discarded as it is unable to distinguish between the real and fake images. There are two limitations in the GAN architecture. Firstly, the Z-space is tied to a pre-defined distribution. This means that whichever distribution is chosen for the Z-space plays a key role in the generation (in fact, the generator is entirely based on this distribution). Moreover, we have no control over the generation process. In other words, we cannot select for the generation of brains with certain attributes such as age, while keeping everything else fixed. Given these limitations, the StyleGAN provides a potentially better baseline architecture for such a desired model and is detailed in the following section.

2.2 StyleGAN

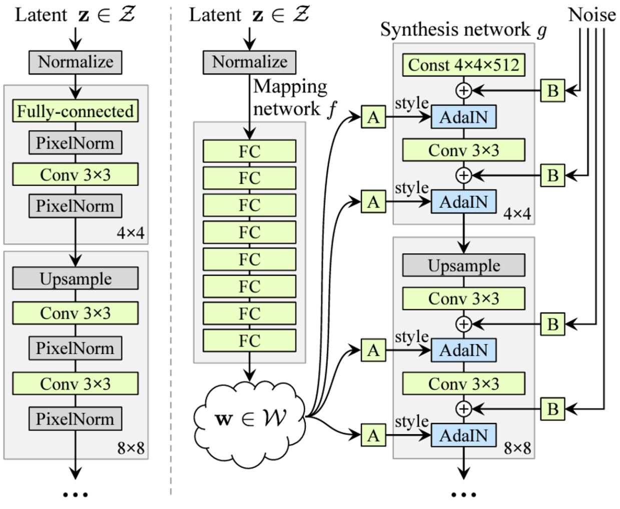
The StyleGAN addresses the limitations of the GAN by modifying the generator architecture, as shown in Figure 2. As can be seen, there are key changes in the StyleGAN architecture. Most notably



$x_{real} \sim P_{real}(x)$
$y_{real} = 1$
$\hat{y} = D(x)$
Is the image real or fake?
Discriminator
z ~ P(z)
Generator
$x_{gen} \sim P_{gen}(x)$
$y_{gen} = 0$

Figure 1: GAN Architecture

for this study, instead of directly using the z in the synthesis network, StyleGAN's generator first maps $z \sim Z$ to another vector, $w \sim W$ via 8 fully connected layers. W space is an intermediate latent space which does not have any pre-defined distribution, instead it is learned during training. It is hypothesized by the authors that there is a pressure during learning for the W space to learn a more disentangled (i.e. linear) representation as it would be easier to generate realistic images from a disentangled representation (8). Thus, the overall idea is that the StyleGAN generator architecture provides a strong baseline for a disentangled latent space where we can theoretically control for different attributes of an image.



(a) Traditional

(b) Style-based generator

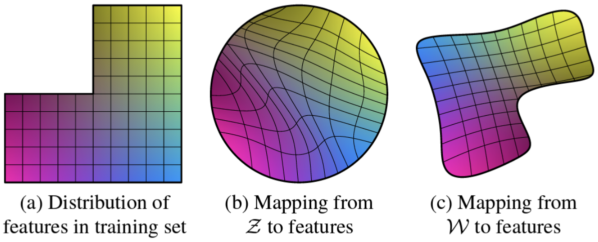
Figure 2: StyleGAN modified generator (8)

2.3 Disentanglement Metrics

Obtaining a disentangled latent space has clear advantages in downstream tasks, including controllable generation, interpretability of results, and reduction of complexity (10). However, aside from visual inspection, it is important to quantify this concept with metrics. Two metrics which are formulated in the StyleGAN paper are described below (8).

2.3.1 Perceptual Path Length

As mentioned, the latent space Z of traditional generative models has a pre-defined distribution and thus is constrained in its ability to be representative of the training data. This is best visualized in Figure 3 as shown in (8). In the case where some combination of features is missing (e.g. young age with many lesions), the mapping from Z to features is forced to become non-linear in order to prevent the sampling of an invalid combination. By adding W, the warping can be undone as W does not rely on a fixed distribution. From this illustration, and from the definition that Z is a Gaussian distribution, it is clear that linear changes in the latent space would result in non-linear changes in the image space. Thus, if we linearly interpolate between two points in Z space, we would not expect to see smooth changes in the image. In terms of disentanglement, one would not be able to interpolate in a highly entangled latent space without seeing drastic changes in the image space. Perceptual path length (PPL) quantifies how drastic these changes in the image space are as interpolation is performed in the latent space. A keypoint when developing PPL is how to define the notion of similarity. Of course, one way is generating the pixel or voxel wise similarity in the image space, but this has clear limitations. Consider an image of a zebra and a soccer ball. Although both are black and white and thus may have high pixel-wise similarity, it is clear to humans that a basketball and soccer ball are closer in similarity than a zebra and soccer ball. Thus image space similarity would not be sufficient. To combat this, (8) suggest using the feature space of the deep learning classifier VGG-16 instead. VGG-16 is a 16 layer neural network that is noted to have a feature space that is highly similar to human perception out of many classification networks (11) (12). Returning to the previous example, if a classifier is trained to classify objects, it is clear that it would classify a soccer ball and basketball similarly, whereas it would classify a soccer ball and zebra differently. We would expect that the final layer of a classifier before classification would have a feature space that represents our notion of similarity, and thus the distance in feature space between similar objects should be reduced as compared to dissimilar objects. This distance corresponds to the perceptual difference between two images.


(a) Distribution of features in training set

(b) Mapping from $\mathcal{Z}$ to features

(c) Mapping from $\mathcal{W}$ to features

Figure 3: Latent Spaces with respect to the distribution of features in the training set (8)

Thus to find the PPL between two images that were generated from two points in latent space, we can subdivide the a latent space interpolation path into linear segments and find the sum of these perceptual differences. Instead of taking the limit of these subdivisions as they approach 0, it is easier in practice to use a small subdivision, such as $\epsilon = 10^-4$. Given the construction of the StyleGAN and the hypotheses mentioned above, we would expect that the mapping to W space would result in a more linear latent space, and thus the PPL metric would decrease. The average perceptual path length (PPL) in W is calculated by

$$PPL = E\left[\frac{1}{\epsilon^2} d\left(G(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t)), G(\mathrm{lerp}(f(\mathbf{z}_1), f(\mathbf{z}_2); t+\epsilon))\right)\right], \tag{2}$$

where $\mathbf{z}_1, \mathbf{z}_2 \sim P(\mathbf{z})$, $t \sim U(0,1)$, G is the generator, and $d(.,.)$ evaluates the perceptual distance between resulting images. If we instead change the equation to

$$PPL = E\left[\frac{1}{\epsilon^2} d\left(G(f(\mathrm{lerp}(\mathbf{z}_1, \mathbf{z}_2; t))), G(f(\mathrm{lerp}(\mathbf{z}_1, \mathbf{z}_2; t+\epsilon)))\right)\right], \tag{3}$$

where we are performing interpolation in Z space before mapping to W space, we would expect that PPL score would increase.

2.3.2 Linear Separability

Another metric as developed by (8) is linear separability. The outline of the metric is shown in Figure 4. Once a StyleGAN is trained, we should be able to generate images that resemble those in the training distribution and thus can be labeled to have the attributes used in the training set. For example, in our dataset each sample has been labeled as male or female. This metric uses the idea that if a latent space is disentangled, then two sets of samples with opposite labels for a binary attribute should be easily separated via a linear hyperplane in the latent space. Thus, by generating images from $\mathbf{z} \sim P(\mathbf{z})$, as W space is supposedly more disentangled, we would expect to be able to easily divide the points in W space. To first generate the labels, an auxiliary classifier network needs to be able to classify the generated images to the labels of the binary attribute. The samples are then sorted according to classifier confidence, and the least confident half of samples is removed. A linear SVM can then be fit to predict he label based on the latent space point and the points divided by this plane can be classified. Finally, the conditional entropy $H(Y|X)$ is calculated, where X are the classes predicted by SVM and Y are the classes determined by the pre-trained classifier. By its definition in information theory, $H(Y|X)$ gives how much additional information is required to determine the true class of a sample, given that we know on which side of the hyperplane it lies. Thus, a low value is desired as it suggests that the latent space is disentangled for the given attribute.

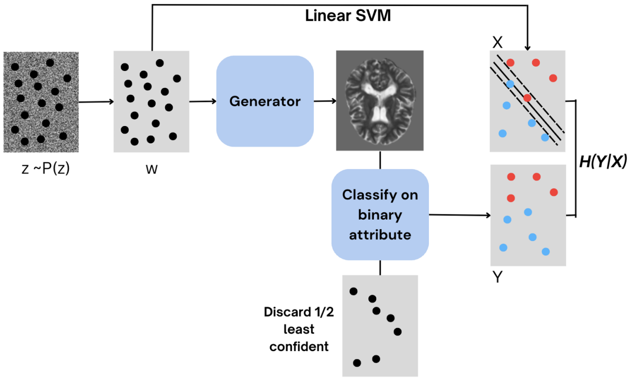


Figure 4: Linear separabiltiy metric formulation

3 Methods

3.1 Resources and Datasets

The dataset used will be the private dataset provided by the Probabilistic Vision Group (PVG). This includes a total of 7542 sample scans from MS patients. The mpMRI scans were acquired at multiple different institutions with different clinical protocols and various scanners. Computation will be done using GPUs available through the Probabilistic Vision Group lab, with Compute Canada clusters also also available if needed.

3.2 VGG-16 Training

Although a pre-trained VGG-16 model is available at here (12), a VGG-16 model using our training data was created to ensure that the feature space corresponds to the attributes that we care about. The results were then compared to the pretrained model and are discussed in 4. The architecture of the VGG-16 model is shown in Figure 5. As can be seen, it contains five convolutional blocks with two convolutional layers followed by a pooling layer. The final block contains 3 fully connected layers before outputting a classification on four heads: age category, disease score, disease type, and sex. During training, the cross-entropy loss function was used as each head was considered categorical. Results on the validation set can be seen in 4.

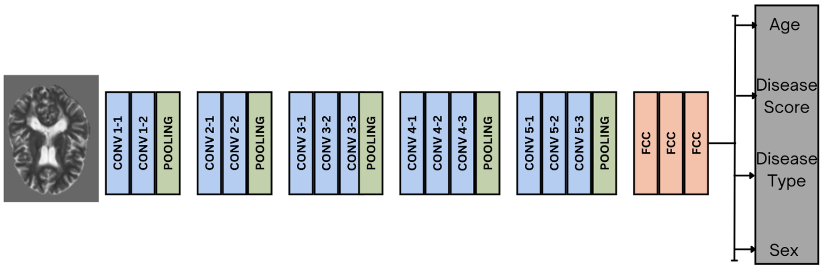


Figure 5: VGG-16 Architecture trained with data attributes

3.3 StyleGAN Training

Although not the focus of the study, a brief review of the StyleGAN training is provided. To create the StyleGAN model, the 3D StyleGAN created in the lab was altered to work with 2D data. The parts of the StyleGAN were initialized as different models, namely the generator and mapping network (from Z to W space) for training, the generator and mapping network for validation, and the discriminator. The training and validation models were separated to ensure no data leakage.

The mapping network consisted of 3 layers instead of the 8 that were used in the StyleGAN paper (8). This was due to the tradeoff between saving computation time and complexity of the mapping. The architecture of the generator was based off of the Synthesis network in Figure 2. The number of blocks varied depending on the number of features in the latent space. As shown in the figure, the size is gradually upsampled in each block. At each block, the information from the W space is added via a learned affine transformation, A and incorporated as the "style" (in the context style transfer literature) through Adaptive Instance Normalization (AdaIn). Additionally, in an attempt to separate stochastic changes of an image from context (this can be thought of as hair placement in an image of a face, for example), noise is injected into every block. The discriminator is the same as in (13) (see the paper for more details). The loss used is the Wasserstein loss with gradient penalty as seen in (14) due to findings in the 3D StyleGAN developed at the lab.

3.4 Alternate Solution: Removal of StyleGAN W - space

We hypothesized that adding the W space in StyleGAN allows for a more disentangled latent space. To test this hypothesis, the architecture was kept the same asides from the mapping network which was removed. Thus, the vectors in Z space were sent directly to the synthesis network. As the samples from Z space follow a Gaussian distribution, it is expected that a well constructed generator that can generate high quality images would not have a disentangled Z space.

3.5 Evaluation

The VGG-16 network was evaluated by its classification performance (accuracy and ROC) for the four heads. To evaluate the disentanglement of the W and Z space in the StyleGAN, and the Z space in the above alternate formulation, two metrics were evaluated: PPL and linear separability. Both of these are described in greater detail in 2.

As discussed, PPL is a metric used to quantify how linear changes in the latent space correspond to changes in the image space. After training the StyleGAN with different latent sizes, two methods for interpolation was used. Firstly, two points were selected from Z space and 150 points were linearly interpolated in between them. These samples from the Z space were then propagated through the mapping network into their corresponding W space representations, and then through the trained StyleGAN generator into their corresponding image representations. Their image representations were then fed into the trained VGG-16 network (we used the pre-trained version from (12) due to the identical trends seen above) to get their feature embedding in the final layer of the network. As described above, their distance in the feature space was then calculated. Similarly, for the second approach, two points were selected from Z space and then mapped to their corresponding points in W space. Once in W space, 150 points were sampled from the linear interpolation between the two points. As before, these points were then put through the generator and finally into a feature embedding in the VGG-16 network. To make an intuitive plot, the distance between each interpolation step in the linear path and the first image was stored. Thus, if we expect a linear path, then the PPL from the first point in the interpolation would be the farthest from the last.

To compare to the baseline of not using a mapping network, and thus restricting the only latent space to a Gaussian distribution, the mapping network was removed from the architecture. This architecture was then trained with different sized latent spaces, and the disentanglement metrics were calculated for it.

4 Results

4.1 VGG-16 Training

The VGG-16 model shown in Figure 5 was trained to obtain a feature space representative of human perception of similarity. The classification scores were monitored on the validation set during training to monitor the learning progress, as shown in Figure 9 of Appendix A. It is notable that compared to sex, age, and MS type classification, the disease score is more difficult to learn from the image. This is confirmed by the best MAE for disease score in Table 1 as compared to age. Finally, the classification for sex and MS Type both achieve >80% accuracy. It is noted that the VGG-16 network was not optimized to achieve the highest classification scores and could be improved. Moreover, the VGG-16 network was used for its feature embedding layer, which has been used in the past due to its representation that is highly similar to human perception (12) (11).

Table 1: Classification Scores in VGG-16 on the Validation set

	Numerical Attributes (MAE)		Categorical Attributes (Accuracy)	
Metric	Age	Disease Score	Sex	MS Type
Best	0.395	0.672	0.877	0.836
Last	0.434	0.716	0.858	0.8

The VGG-16 model trained on our data was compared to the pre-trained VGG-16 model available by the authors of (12) during the training of the StyleGAN. This was to ensure that the changes seen to the PPL score during training followed similar trends. As can be seen in Figure 8, both the pre-trained and data-specific VGG-16 model follow the same trends for varying latent space sizes. This ensures that the VGG-16 training was done properly. It is hypothesized that before training, the generator is generating images that are very similar to each other in the image space as it has not learned to generate real images yet. Thus when sent to a feature embedding, these images are sent to the same or very similar representation. As training continues, the generator is able to create better quality images, and as the quality of the image increases, the differences in image space increase. Thus, when the images are sent to the feature embedding of the VGG-16 network, there is greater difference in their feature embedding.

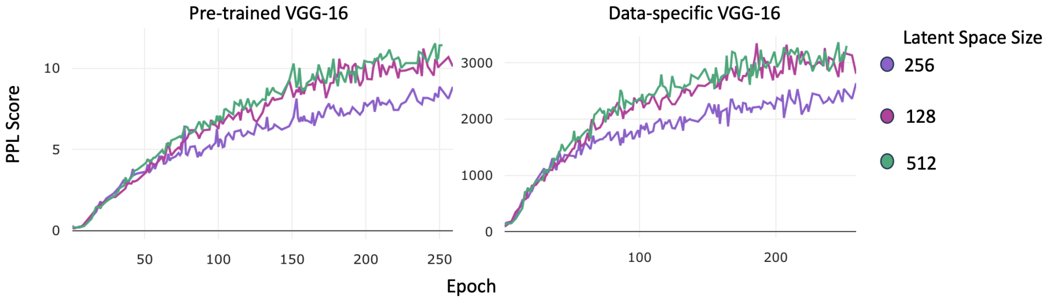


Figure 6: PPL Score during training of StyleGAN using pre-trained and data-specific VGG-16

4.2 Disentanglement Metrics

4.2.1 Perceptual Path Length

The results of the PPL metric are shown in Figure 7. As can be seen, with a larger latent space, the interpolation whether done in Z or W space looks similar. Notably, in the latent size of 512, the interpolations look nearly identical. Given that points from the Z space are randomly sampled from a Gaussian distribution before being mapped to the W space, it is impossible that one could linearly interpolate in a Gaussian space and see linear changes in the feature embedding of the image space (or the image space itself). Thus, due to the large latent size, it is possible that the model is over-parameterized and is learning a linear mapping from a non-linear space. This hypothesis is confirmed when testing other latent sizes. As the latent space size decreases, the points linearly sampled from Z space no longer correspond to linear changes in the feature embedding of their generated images. In the corresponding GIFs of the interpolated points in image space, [1], one can clearly see the difference between the interpolations from Z and W space. With smaller latent sizes, changes in the image space from Z interpolation appears to be "faster/slower" in some regions. This theoretically makes sense as the Z space should be curved. From the interpolations done in W space, the changes in image space appear to be very linear, with no drastic changes from one step to the next. This implies a linear latent space, as desired.

Given the hypothesis about the W space, the mapping network was completely removed. It was hypothesized that by removing it, we would not be able to map linear changes in Z space to linear changes in the image space or its feature embedding space. Latent space sizes of 128 and 32 were tested, and it can be seen that the GAN style architecture is able to learn linear mappings from linear interpolations in Z space to interpolation in the feature embedding of the image space. This is confirmed in the interpolation in image space of the corresponding latent vectors, where the transitions are smooth (as seen in the attached file). In Figure 8, the interpolation in W for the StyleGAN is compared to the interpolation in Z for a GAN. It can be seen that they both follow linear trends. This was a surprising result given the hypotheses of the work. Although the points in the latent space are sampled from a Guassian distribution, the network is able to understand linear changes in the latent space and map them to linear changes in the image space. This may point to the model's capacity to learn a linear mapping from a non-linear distribution.

Moreover, after visual inspection of the images generated (and in the GIFs attached), it can be seen that the image quality does not significantly degrade with decreasing latent size. Thus, the same quality image that can be represented with a 512 latent size can also be represented with a 32 latent size for both the StyleGAN and modified architectures. This leads to the conclusion that the generator architecture is not optimized, as we would expect the possibility of higher quality images with higher dimension latent spaces. Although the model appears to be able to learn a linear mapping from either the Z space in a GAN or the W space in a StyleGAN, the quality of the images has not been optimized.

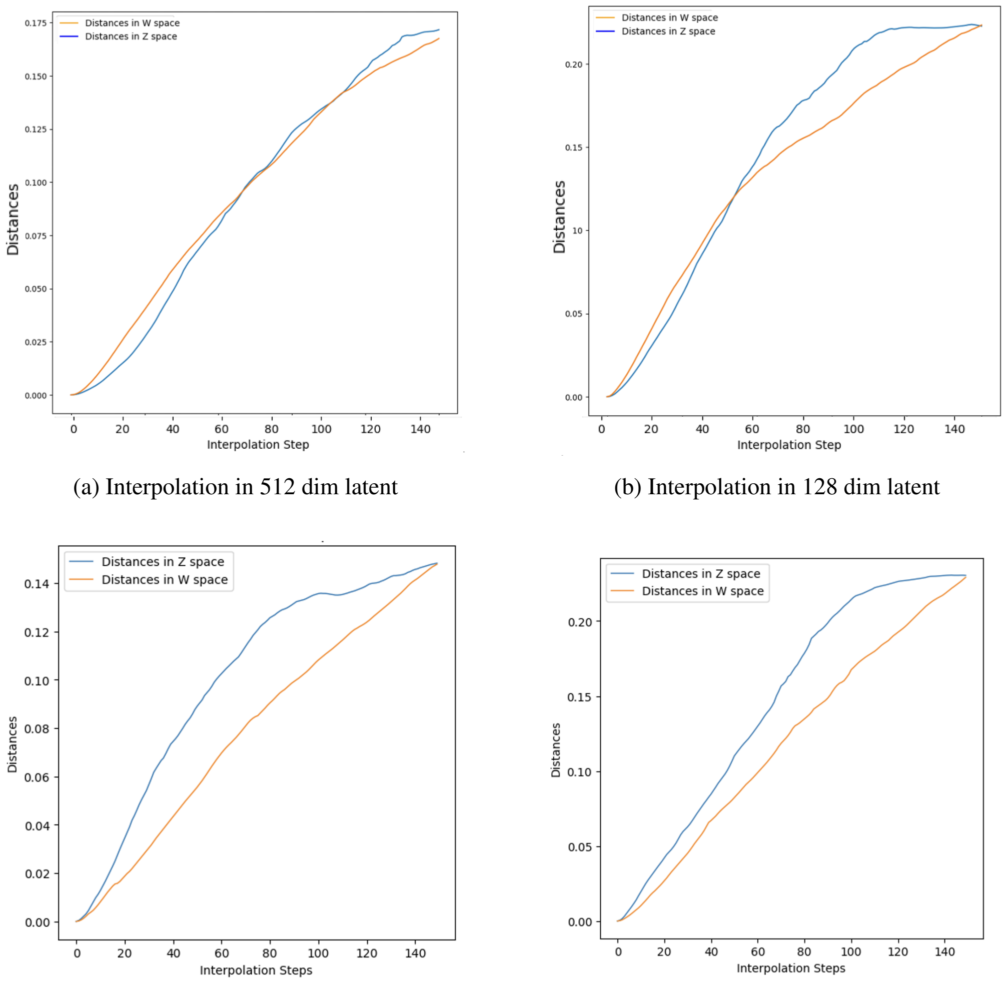
4.2.2 Linear Separability

Additionally, the linear separability metric, which is the conditional entropy $H(Y|X)$ where X are the classes predicted by SVM and Y are the classes determined by the pre-trained classifier, was calculated as well as the accuracy of the SVM in the latent space. These results can be seen in the table below. Notably, the Z space in a GAN can be separated such that a hyperplane can separate the two classes male and female. This indicates that the features in the latent space that corresponding to the sex of the patient are easily linearly separable even in a Gaussian distribution. In fact, with a latent space of size 32, the accuracy of the SVM in the Z space of a GAN is on the line of the accuracy of the SVM in the W space of a StyleGAN.

5 Future Work and Conclusions

In this work, we have developed and trained a 2D StyleGAN with various latent sizes and have measured the disentanglement of the latent space. We did the same with a variant of the StyleGAN without the mapping network, more similar to traditional GAN architectures. From the results,

[1] attached in .ppt file in submission



(a) Interpolation in 512 dim latent

(b) Interpolation in 128 dim latent

(c) Interpolation in 64 dim latent

(d) Interpolation in 32 dim latent

Figure 7: Linear Interpolations in Z and W space and the corresponding PPL

Table 2: Conditional Entropy $H(Y \vert X)$ and Accuracy of SVM classifier

	Conditional Entropy $H(Y \vert X)$		Accuracy of SVM
Latent Dim	Model, Space	Value	Value
64	StyleGAN, W space	0.324 ± .024	0.962 ± .006
	StyleGAN, Z space	0.375 ± .041	0.948 ± .013
32	StyleGAN, W space	0.384 ± .034	0.949 ± .008
	StyleGAN, Z space	0.465 ± .059	0.915 ± .019
	GAN, Z space	0.387 ± .035	0.947 ± .008

although the intermediate latent space in the StyleGAN appears to be sufficiently disentangled, removing the mapping network has similar results in terms of disentanglement metrics and image quality. Moreover, the resulting image quality speaks to the trade-off between disentanglement and high quality generation (7). This leads to the conclusion that the generator architecture has not been optimized and thus needs to be the next step in order to properly assess the disentanglement metrics, and the potential benefits of the W space in the StyleGAN architecture. Due to this limitation, more work needs to be done on creating an optimized generator. Of course, with higher image quality, there is a possibility that the network will not be able to learn linear mappings from the space. Once

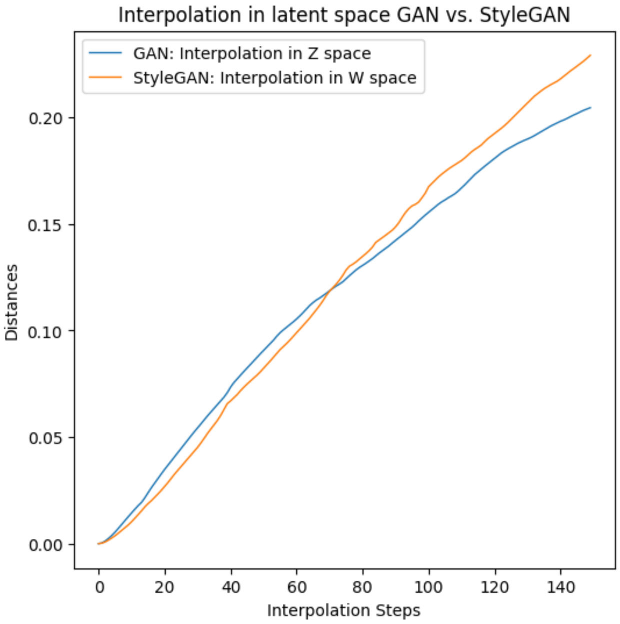


Figure 8: PPL between feature embeddings using the GAN and StyleGAN networks

the generator is optimized, we can better understand the trade-off between a disentangled latent space and and high image quality generation.

With a disentangled latent space, the future research directions of this work would be to condition the model on certain attributes. Once it is understood which features in the latent representation correspond to which attributes, altering specific attributes can easily be done in the latent space by changes to their features in the latent space (4). By creating such a conditional model, one would be able to have a controlled generation of images - a requirement if such models are to be implemented in the medical field. By creating tools to analyze the disentanglemnt of the latent space this work is an exciting step towards creating controllable generative models in the context of medical imaging.

References

[1] M. J. P. M. M. R. R. A. F. K. J. M. G. J. H. M. P. S. N. U. R. A. Afshin Shoeibi, Marjane Khodatars, "Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review," *Computers in Biology and Medicine*, 2021.

[2] R. Jason N. Itri, Rafel R. Tappouni, "Fundamentals of diagnostic error in imaging," *Radiographics*, 2018.

[3] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable? insights imaging," *Insights Imaging*, 2018.

[4] M. K. S. S. C. Zhenliang He, Wangmeng Zuo, "Attgan: Facial attribute editing by only changing what you want," *IEE Transactions on Image Processing*, 2018.

[5] S. O. Mehdi Mriza, "Conditional generative adversarial nets," *arXiv*, 2014.

[6] S. A. T. Tian Xia, Agisilaos Chartsias, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Medical Image Analysis*, 2020.

[7] Y. W. W. Z. Xuanchi Ren, Yang Tao, "Learning disentangled representations by exploiting pretrained generative models: a contrastive learning view," *ICLR*, 2014.

[8] T. A. Tero Karras, Samuli Laine, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[9] M. M. B. X. D. W.-F. S. O. A. C. Y. B. Ian J. Goodfellow, Jean Pouget-Abadie, "Generative adversarial networks," *arXiv*, 2014.

[10] J. B. G. G. Marc-Andrew Carbonneau, Julian Zaidi, "Measuring disentanglement: A review of metrics," *EEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] S. A. RT Pramod, "Improving machine vision using human perceptual representations: The case of planar reflection symmetry for object classification," *EEE Transactions on Neural Networks and Learning Systems*, 2020.

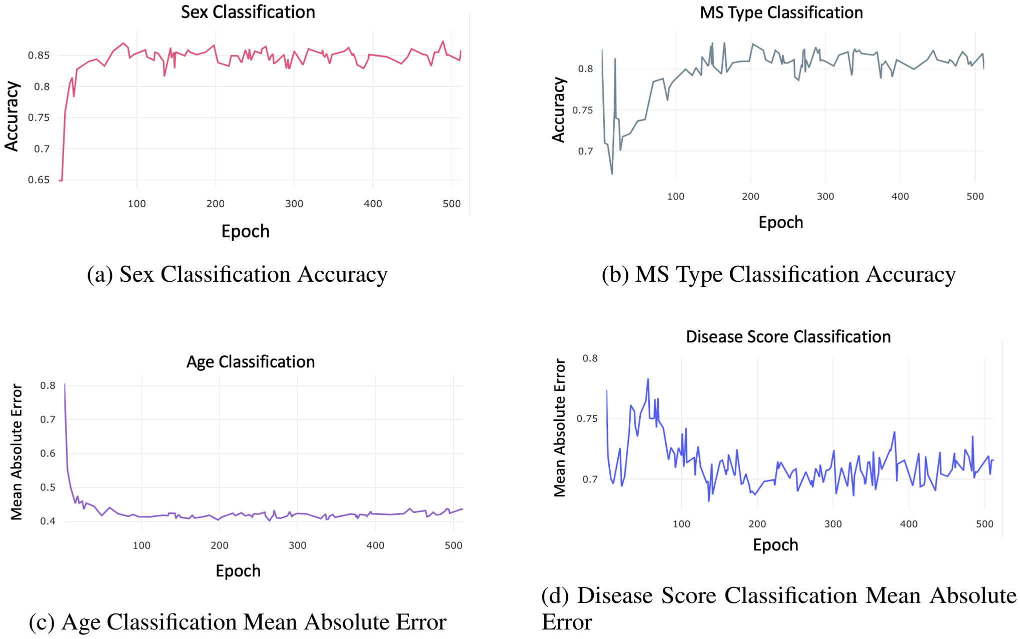
[12] A. A. E. E. S. O. W. Richard Zhang, Phillip Isola, "The unreasonable effectiveness of deep features as a perceptual metric," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[13] S. L. J. L. Tero Karras, Timo Aila, "Progressive growing of gans for improved quality, stability, and variation," *arXiv*, 2017.

[14] M. A. V. D. Ishaan Gulrajani, Faruk Ahmed and A. C. Courville, "Improved training of wasserstein gan," *Advances in neural information processing systems*, 2017.

A Appendix

Additional results.



(a) Sex Classification Accuracy

(b) MS Type Classification Accuracy

(c) Age Classification Mean Absolute Error

(d) Disease Score Classification Mean Absolute Error

Figure 9: Classification Performance on Validation Set