

# Latent space manipulation for high-resolution medical image synthesis via the StyleGAN

Lukas Fetty<sup>a,\*</sup>, Mikael Bylund<sup>b</sup>, Peter Kuess<sup>a</sup>, Gerd Heilemann<sup>a</sup>, Tufve Nyholm<sup>b</sup>, Dietmar Georg<sup>a</sup>, Tommy Löfstedt<sup>b</sup>

<sup>a</sup> Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria

<sup>b</sup> Department of Radiation Sciences, Umeå University, Umeå, Sweden

Received 20 December 2019; accepted 1 May 2020

## Abstract

**Introduction:** This paper explores the potential of the StyleGAN model as an high-resolution image generator for synthetic medical images. The possibility to generate sample patient images of different modalities can be helpful for training deep learning algorithms as e.g. a data augmentation technique.

**Methods:** The StyleGAN model was trained on Computed Tomography (CT) and T2-weighted Magnetic Resonance (MR) images from 100 patients with pelvic malignancies. The resulting model was investigated with regards to three features: Image Modality, Sex, and Longitudinal Slice Position. Further, the style transfer feature of the StyleGAN was used to move images between the modalities. The root-mean-square error (RMSE) and the Mean Absolute Error (MAE) were used to quantify errors for MR and CT, respectively.

**Results:** We demonstrate how these features can be transformed by manipulating the latent style vectors, and attempt to quantify how the errors change as we move through the latent style space. The best results were achieved by using the style transfer feature of the StyleGAN (58.7 HU MAE for MR to CT and 0.339 RMSE for CT to MR). Slices below and above an initial central slice can be predicted with an error below 75 HU MAE and 0.3 RMSE within 4 cm for CT and MR, respectively.

**Discussion:** The StyleGAN is a promising model to use for generating synthetic medical images for MR and CT modalities as well as for 3D volumes.

**Keywords:** StyleGAN, Image synthesis, Latent space

## Introduction

Medical imaging has been highlighted as one of the areas where deep learning has the largest implications, and greatest potential [1–3]. For instance, decision support systems for radiological evaluations of images have been described in several recent reviews and publications as an area where deep learning can lead to significant benefits for the patients [4]. There are many other potential applications, such as segmentation and delineation of organs or even tumour regions, which

are important for radiological and radiotherapy applications [5,6], image improvement and super-resolution [7], creation of attenuation maps for attenuation correction of PET/MR data or for novel individualised radiotherapy treatment planning concepts [8–10], etc.

These applications all require, and will continue to need, large sets of training data that span the population variability well in order to avoid overfitting and to produce reliable results. This is problematic for medical applications since medical data is usually scarce, and it is challenging to share

\* Corresponding author: Lukas Fetty, Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria.  
E-mail: [lukas.fetty@meduniwien.ac.at](mailto:lukas.fetty@meduniwien.ac.at) (L. Fetty).

medical data with third parties or even between different hospitals because of patient integrity concerns [1].

Since generative adversarial networks (GANs) were introduced in 2014 [11], they have been used for image augmentation in numerous medical applications [12,13]. GANs are also used in many other applications [14], such as in image registration, image segmentation, and image-to-image translation tasks, just to name a few.

One of the main reasons behind the development of the GAN was the possibility to synthesise artificial data from completely unlabelled training data. This application has also been investigated by several research groups [15,16]. However, the methods that have been available for generating synthetic images have not scaled to high-resolution images. Because of this, the publications that describe the use of synthetic images have mostly dealt with the low-resolution case, and have thus been severely limited.

Recently, Karras *et al.* [17] proposed a novel GAN model that includes a progressive increase of the output image resolution during training, and the end result is the successful synthesis of realistic high-resolution images. They then further improved the model and also made it possible to include stochasticity and style transfer in the generation process; the improved progressive GAN was called *StyleGAN* [18]. The output images are now of sufficiently high resolution and quality that the generated images could potentially be used to augment a medical image dataset. The ability to train on synthetic images would thus alleviate the small dataset problem, allowing deep learning models to be trained on large amounts of synthetic data.

Since the latent space becomes a high-level representation of the images, certain known attributes of the images and the corresponding latent vectors can be used to learn a function or direction that describes the attributes [19,20]. In order to use the generated images for training *e.g.* deep learning models, it is imperative to understand the latent space, what it can encode, and how it is organised.

In the present study, the StyleGAN model's latent style space was investigated, providing a deeper understanding of the internal structure of the network. Further, the possibilities to manipulate the latent space was reexamined in order to generate customised high-dimensional medical images. The StyleGAN model was investigated after being trained on images of two different modalities:  $T_2$ -weighted magnetic resonance (MR) images and computed tomography (CT) images, captured in the pelvic regions of both male and female patients with various cancer diagnoses. Three image attributes (Image Modality, Sex, and Longitudinal Slice Position) were selected and methods were developed to manipulate them such that images with custom representation of these attributes could be generated in a controlled manner.

## Material and Methods

### Data

The data used in this study were collected from 117 patients undergoing treatment at Umeå University Hospital. The patients were mainly diagnosed with either prostate, rectal, or gynaecological cancer, and were imaged using both MR and CT as part of the routine clinical workflow. The MR images were acquired with a GE 3T SIGNA PET/MR, and the CT images with a Philips Brilliance Big Bore. See the [supplementary material](#) for acquisition details. The data collection was performed according to an existing ethical permit (number 2018-234-31 M), and informed consent was obtained from each patient. Patients with metal hip implants were excluded, leaving 15 female (mean age 68 years) and 85 male patients (mean age 70 years). The MR images were  $T_2$ -weighted and were bias corrected using N4ITK; the CT images were rigidly registered to the MR images using Elastix with standard settings and were further resampled to the same matrix dimension of  $512 \times 512$  pixels. The preprocessing was performed using MICE Toolkit (NONPI Medical AB, Umeå, Sweden; [www.micetoolkit.com](http://www.micetoolkit.com)). The MR images were normalised by scaling the range [0, 2500] to [-1, 1], and the CT images by scaling [-1024, 1500] to [-1, 1], without truncation. In total 17,542 images were used for training which included on average 88 images per patient and modality.

### Network architecture

The StyleGAN architecture differs from the original GAN model in several ways. As presented in Karras *et al.*, the StyleGAN model differs in three major ways:

- The main network is a progressively growing GAN where the generator network first learns to generate low-resolution images, and then progressively generates larger and larger images as the training progresses. This makes the network converge even for high-dimensional outputs.
- The input to the network is a  $d$ -dimensional independent Gaussian random vector with zero mean and variance one. This input space, denoted  $Z$ , is mapped to an intermediate latent space, denoted  $W$ , by a dense neural network. This step transforms the input space to a *style* space, that is assumed to be more disentangled, having different image features encoded along different (approximately) orthogonal dimensions in the style space. Smooth interpolations where individual features are controlled should therefore be possible in the  $W$  space. The  $W$  space is then normalised (adaptive instance normalisation, AdaIN) [21], and fed to the different layers of the generator network.

- Finally, the StyleGAN model also accepts noise injections at each resolution in order to introduce stochastic details in the images.

### Training

The StyleGAN model<sup>1</sup> was implemented in PyTorch v1.0 [22] in such a way that it takes a random latent vector and outputs a 2D image. The Adam optimiser [23] was used with momentum parameters  $\beta_1 = 0$  and  $\beta_2 = 0.99$ . The weights were exponentially averaged with a decay rate of 0.999 as in [17]. The initial learning rate was set to 0.001 for both the generator and the discriminator, and to 0.0001 for the mapping network that transforms the latent vectors from the Z to the W space. Mixing regularisation was also included, which injects a second latent vector at a random resolution during training. As in the original StyleGAN paper, two loss functions were included, namely the Wasserstein loss with a gradient penalty (WGAN-GP) [24] and a non-saturating loss [11] with  $R_1$  regularisation [25]. The mini-batch size was progressively decreased from 256 to 8 images per batch, and was decreased when the resolution was increased.

The StyleGAN model was trained on two NVIDIA RTX 2080 Ti GPUs for up to a total of 2,400,000 gradient updates.

### Image quality

The Fréchet inception distance (FID) was used to evaluate the quality of the generated images. The FID was computed using all the training images and using 10,000 random images generated with a fixed truncation level of 0.7 [26].

The FID was computed at every 100,000 updates from 700,000 to 2,400,000 and used to select the final model.

### Latent space manipulation

The manipulation of the modality (MR to CT and the reverse), the sex of the patient, and how to change the longitudinal slice position of the patient (the slice index in the image volume) was investigated.

In order to find latent directions that encode the features, an image encoder-decoder was constructed to generate style vectors  $w$  from images and vice versa. It was built using the StyleGAN generator as a (fixed) decoder. The encoder had the same structure as the StyleGAN generator, but in reverse, and with instance normalisation [27] instead of AdaIN. The convolutional part of the encoder had an output resolution of  $4 \times 4 \times 512$ , and this hidden representation was fed to a four-layer dense neural network with LeakyReLU activations with a negative slope of 0.2, resulting in a final 512-dimensional output in the W space of the StyleGAN.

The encoder network was trained using the RAdam optimiser [28] with  $\beta_1 = 0$  and  $\beta_2 = 0.99$ , as for the StyleGAN training. The weights were exponentially averaged with a decay rate of 0.999 [17]. The loss function was the sum of the Euclidean distance between the input random W space vectors and the reconstructed latent style vectors output from the encoder and an  $\ell_1$  loss between the input image and a StyleGAN regenerated image. The encoder network was trained on 224,000 randomly generated images and corresponding style vectors with a mini-batch size of 8. The random images were generated using a truncation level of 0.7.

The encoder was then used in a refinement process to embed the training images into the W space. The generated style vector was applied as an initial guess and was further refined by minimising the loss between input image and the regenerated StyleGAN image. Since the initial latent vector output of the encoder can be unstable, the refinement was introduced to reduce differences between the target image and the generated image. The loss function for the refinement was a combination of feature and  $\ell_2$  loss,

$$L_{\text{refine}}(w) = L_{\text{features}}(G(w), x) + \|G(w) - x\|_2 \quad (1)$$

where  $x$  is the input image and  $G(w)$  is the image that the StyleGAN regenerated from the style vector  $w$ . The feature loss is defined as:

$$L_{\text{features}}(w) = \sum_{r \in R} \|D_r(G(w)) - D_r(x)\|_2 \quad (2)$$

where  $D_r$  is the output of the first convolution layer of the discriminator at resolutions  $r \in R = \{64, 128, 256, 512\}$ . RAdam was used with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , with an initial learning rate of 0.001. The number of iterations was limited to 350 to constrain the refined latent vectors to lie near the initial guesses.

The refined style vectors from the training images were finally used to model the latent space for feature manipulation. The overall workflow can be seen in Figure 1.

### Latent space models, and model selection

The manipulations were performed by creating prediction models for the features of interest, for which the ground truth was known from the training data. For this, logistic regression was used as a baseline model, and hyper-parameter searches were performed in order to attempt to find suitable deep dense neural networks for predicting the features better than the baseline logistic regression models.

In order to manipulate the latent style vectors, the representation of each style vector in the last hidden layer of the found prediction model had to be transferred back to the latent style space of the StyleGAN. Therefore, another hyper-parameter search over dense neural networks was performed to predict the corresponding latent style vectors from a vector in the

<sup>1</sup> <https://github.com/roinality/style-based-gan-pytorch>; This repository includes an implementation of the StyleGAN model which was used and adapted for our experiments.

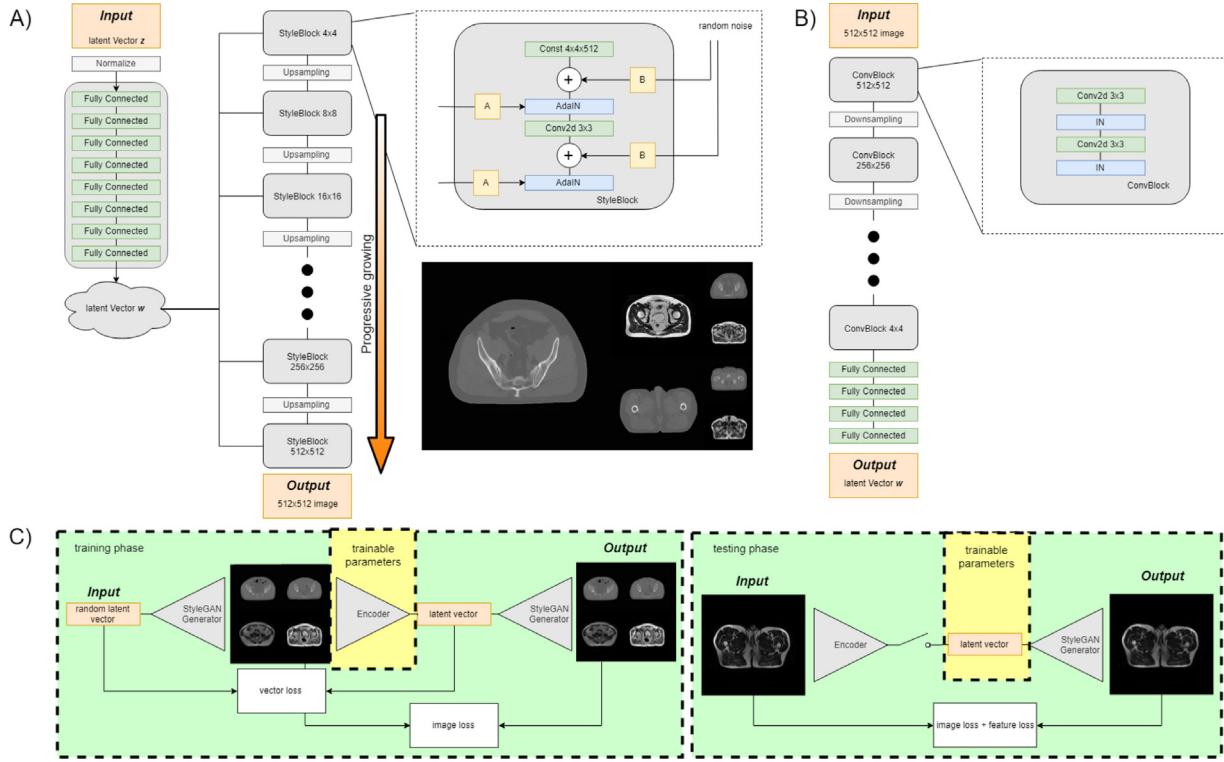


Figure 1. An illustration of the network architectures of the models used in this study. A) An illustration of the architecture of the StyleGAN model, and B) an illustration of the encoder structure used to transform an image to the latent style space. C) An illustration of the training procedure of the encoder and the refinement procedure. The four different images in the training process shown in C) symbolises the batch size. The switch in the testing process in C) defines the injection of the initial guess of the Encoder which is only used in the first iteration.

last hidden layer of the prediction model (denoted the reverse model).

Details on these hyper-parameter searches can be found in the [supplementary material](#).

To determine a direction in the latent space encoding the longitudinal direction, regression networks were trained to predict the normalised slice index (the most inferior slice was set to zero, the most superior slice was set to one, and the other slices were linearly assigned a value between zero and one), but this approach had problems converging, and generally did not perform well. Therefore, slices with a normalised index below 0.4 were instead assigned to class zero (inferior), and slices with normalised index above 0.6 to class one (superior), and classification networks were trained. Hence, the procedure here was the same as that for modality and sex.

In case the validation  $R^2$  (coefficient of determination) for the reverse model times the validation accuracy of the prediction model was lower than the logistic regression baseline validation accuracy, the logistic regression models were selected instead of the dense networks.

#### Interpolating the latent space

The found prediction models were used to manipulate a latent style vector by the transformation

$$\tilde{w} = \text{Reverse}((\bar{w}^* + \tau(\text{Forward}(w) - \bar{w}^*)) + \alpha w_{LR}) \quad (3)$$

where Forward transforms the style vector to the last hidden layer of the prediction model, and Reverse transforms a vector in the last hidden layer of the prediction model back to the style space,  $W$ . The  $w$  is a latent style vector,  $\tau \in (0, 1)$  the truncation level,  $\bar{w}^* = \text{Forward}(\bar{w})$  is the mean latent vector (the mean of 1,000 randomly generated latent vectors) transformed to the last hidden layer of the prediction model,  $\alpha \in \mathbb{R}$  is a weight coefficient for the attribute manipulation, and  $w_{LR}$  the parameter vector in the last hidden layer of the prediction model, *i.e.* in all cases a logistic regression coefficient vector in the last hidden layer of the prediction model (note that the deep dense networks also perform logistic regression in the last hidden layer). The coefficient vector  $w_{LR}$  describes the direction encoding a particular feature, and hence the

direction in which the style vector should be moved in order to change the corresponding feature.

The style space manipulation was investigated with regards to the modality and the longitudinal slice position, since the ground truths were available in those cases from the patient data set.

**Modality** For the modality, 1,000 slices were randomly selected from the initial patient training set and transformed using Equation 3 by moving the corresponding style vector in the direction of the other modality by scaling the decision plane normal by a factor in the range [1,100] in 31 steps on a log scale. Images generated at each of the 31 points were compared to the ground truth image of the other modality. For the generated CT images the mean absolute errors (MAEs) were computed between the true images and the generated images, and for the generated MR images the root mean squared errors (RMSEs) were computed.

**Longitudinal Slice Position** For the longitudinal slice position, all slices from all 100 patients were used. One of the most central 51 slices was selected to be the main slice, and another of the central 51 slices was selected as the sought slice. The main slice was transformed by moving the corresponding style vector in the inferior-superior direction, as identified by the prediction model, by a factor in the range [-0.8, 0.8] and the sought slice was compared to the decoded image corresponding to the transformed main slice by the MAE for generated CT images, and the RMSE for generated MR images.

#### Style transfer

The StyleGAN also allows the generator output to be manipulated by using the style transfer capability of the network, where two latent vectors are included into the generation process (corresponding to source and target images). Style vectors can be injected into the AdaIN to give the network the ability to fuse different representations of the image, such as *e.g.* an MR and a CT image.

The degree of mixing is changed as a function of the injection location, *i.e.* into which resolution layer the second style vector is injected. If the vector is injected into low-resolution layers (*e.g.* 4–64 pixels), this results in strong mixing of the image characteristics. If the vector is instead injected into high-resolution layers (*e.g.* 64–512 pixels), it is mainly colour adaptation of the images that is achieved. This is also seen in the original StyleGAN paper [18].

Style transfer was performed by injecting 1,000 random style vectors from the training data into all seven layers (resolutions of 4–512 pixels), comparing to the paired image from the other modality. The generated CT images were compared using MAE, and the generated MR images using RMSE. Qualitative tests were also performed by visually evaluating the network's style transfer ability.

#### Analysis of the decision surface

In order to improve the understanding of the decision surfaces, and thus the disentanglement of the W space, the

average curvatures of the prediction model's decision surfaces were computed as a means to understand the feature separation in the latent style space. This was done by randomly selecting points on the decision surface, randomly selecting tangent directions, computing the curvature of the graph induced by the intersection of the tangent-normal plane and the decision surface. For each point, the mean curvature at the point was estimated as the mean of 512 random tangent directions, and the mean and standard deviation were computed from 1,000 random points on the decision surface. See the [supplementary materials](#) for details on this procedure.

## Results

The network training took over one month using all the 17,542 images. The image quality increased progressively during training, when using the WGAN-GP loss function. Initial tests also included the  $R_1$  loss function, but led to poor image quality and a failure to converge to meaningful outputs. After 1.2 M updates, the network performance was still increasing, and so the training was continued for another 1.2 M updates.

#### Image quality

The FID score decreased over the course of training. After about 1 M updates the FID score was about 20, and it decreased to just above 12 at 2.2 M updates. A plateau was then observed with FID scores around 12–13 until all 2.4 M updates had been made.

The model at 2.2 M updates had the lowest FID score (with a score of about 12.3). The 2.2 M model was therefore the model that was used throughout this work.

[Figure 2](#) illustrates sixteen random example images generated using the 2.2 M network. See the [supplementary material](#) for more random example images.

#### Model selection

##### Modality

For separating style vectors into those encoding MR and those encoding CT images, the best model ended up being a network with no hidden layers, *i.e.* a logistic regression model. The validation set accuracy was about 1.0.

Since the best model was already a logistic regression model, no reverse models were evaluated for the modality.

##### Longitudinal slice position

To determine a direction in the latent space that encodes the longitudinal direction, the best model had one hidden layer, with 128 neurons in the hidden layer. The dropout rate was about 0.095, and the initial learning rate was about 0.0044.

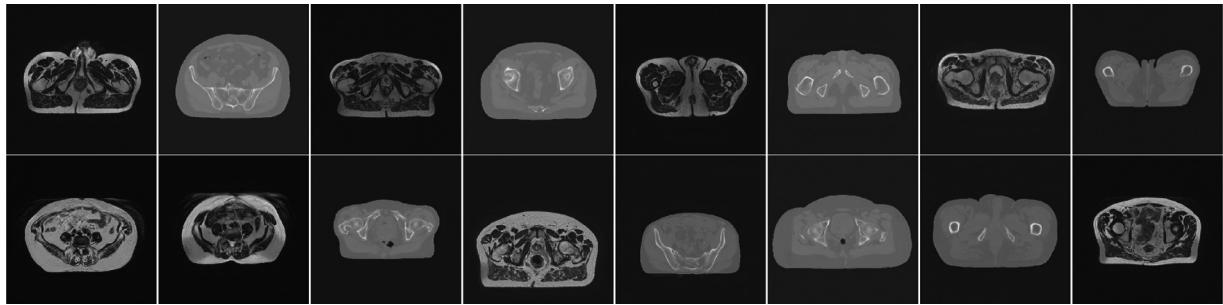


Figure 2. Sixteen random samples from the StyleGAN model trained on 100 patient volume images of paired MR and CT images.

The network was trained for 150 epochs with a mini-batch size of 32. The validation set accuracy was about 1.00.

The best reverse model was a network with three hidden layers, with 300, 152, and 370 neurons in the hidden layers, respectively. The dropout rate was 0.0 (*i.e.*, no dropout was used), and the initial learning rate was about  $6.0 \cdot 10^{-5}$ . The network was trained for 82 epochs using mini-batches of twelve hidden layer vectors. The validation set  $R^2$  was about 0.98. The baseline logistic regression model had a validation set accuracy of about 0.99. Hence, since  $0.99 \geq 1.00 \cdot 0.98$ , the baseline logistic regression model was used for manipulating the patient longitudinal direction feature.

#### Sex

For the model to classify the patient's sex, the best model was a network with two hidden layers, with 128 neurons in the first hidden layer and 54 neurons in the second hidden layer. The dropout rate was zero, *i.e.* no dropout, and the initial learning rate was 0.0013. The network would be trained for 150 epochs using mini-batches of 45 style vectors. The validation set accuracy was about 1.00.

The best reverse model was a network with four hidden layers, with 512, 500, 341, and 263 neurons in the first through fourth hidden layers, respectively. The dropout rate was 0.0, *i.e.* no dropout was used, and the initial learning rate was about 0.00026. The network was trained for 89 epochs using mini-batches of eleven hidden layer vectors. The validation set  $R^2$  was about 0.98. The baseline logistic regression model had a validation set accuracy of about 0.98. Since  $0.98 \geq 1.00 \cdot 0.98$ , the baseline logistic regression model was thus selected for manipulating the patient sex feature.

#### Latent space manipulation

##### Manipulating the modality

The least mean MAE over 1,000 random patient slices was about 73.6 HU when transforming from MR to CT. The least RMSE over the same 1,000 random slices was about 0.35 when transforming from CT to MR (computed on images still in the range  $[-1, 1]$ ). Figure 3 illustrates the errors over the 1,000 random slices when transforming CT to MR images

(left) and correspondingly MR images to CT images (right) by moving from random points from one modality into the domain of the other modality. The error bars correspond to approximate 95% confidence intervals of the means. The least errors thus best correspond to the associated ground truth images.

##### Manipulating the longitudinal slice position

Figure 4 illustrates the manipulation of the longitudinal slice position when moving along the decision surface normal “searching” for a slice at different offsets from the main slice. The top row contains the results for MR images and the bottom row the results for the CT images. The left part illustrates the average errors (averaged over the 100 patients) when interpolating between pairwise slices. The right part illustrates the average errors when interpolating between the centre slice and slices offset from the centre slice, together with the errors at translations using factors in the range  $[-0.8, 0.8]$ . The average distance between slices in the W space was about 0.031 for the MR images and about 0.029 for the CT images (averaged over all patients), in normalised units along the decision surface normal. *I.e.*, to obtain the next inferior slice given a particular slice, we move in the negative direction of the normal a distance of about 0.031 or 0.029, and to obtain the next superior slice given the same slice, we move in the direction of the normal a distance of about 0.031 or 0.029. This worked well starting from any slice, but the errors clearly increase with the distance between the slice positions.

#### Style transfer

The mean errors over the 1,000 random slices were 68.6, 58.6, 58.7, 58.9, 60.4, 459.1, 461.4, and 460.6, for MR to CT transformation when injecting in the zeroth through seventh locations ( $4 \times 4$  through  $512 \times 512$ ), respectively. For the CT to MR transformation, the corresponding RMSE values were 0.344, 0.340, 0.340, 0.340, 0.339, 0.380, 0.377, and 0.377. Figure 5 illustrates an image transfer when the target information is injected in different locations of the network. Injections in the first layer ( $4 \times 4$ ) resulted in general feature mixing like modality and longitudinal slice position. Injecting the images in the middle layers ( $8 \times 8$  through  $64 \times 64$ )

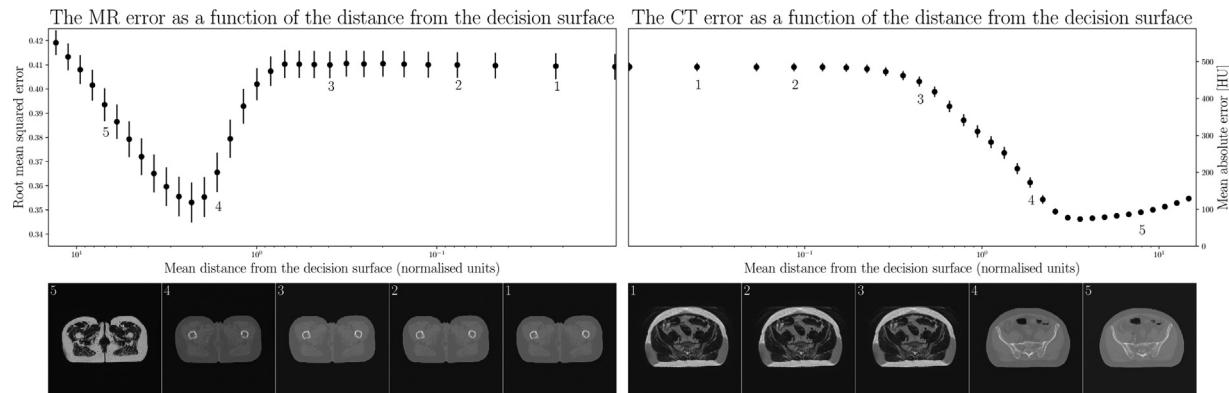


Figure 3. An illustration of moving from the CT domain to the MR domain (left) and vice versa (right). The points are the RMSE and MAE, respectively, over 1,000 random slices, and the error bars are approximate 95% confidence intervals of the means. The images in the lower part of the figure are from a random patient (one random in the left part and another random in the right part), decoded at their corresponding positions on the first axis of the plots. Note the direction of the distance axis in the left plot. The numbers in the image indicate the loss of the images in the lower parts. The numbering counts from one to five, where one is closest to the decision boundary and increases together with the distance to the decision boundary.

changed the modality information of the source to the target one. No changes were observed when the target vector was injected in later layers ( $128 \times 128$  through  $512 \times 512$ ).

### Analysis of the decision surface

See the [supplementary material](#) for illustrations of the decision surfaces of the different models. The curvatures across the surface for the sex had a mean of about -0.0013 with a standard deviation of about 0.0053, the curvatures for the longitudinal slice position had a mean of about -0.00039 with a standard deviation of about 0.00077. Hence, the curvatures were mostly zero, and did not significantly deviate from zero. However, the distribution of average curvatures appeared not to be normally distributed ( $p < 4 \cdot 10^{-80}$  for sex and  $p < 1 \cdot 10^{-28}$  for longitudinal slice position, using the D'Agostino and Pearson normality test), which would be the expected outcome had the average curvatures come from the same distribution.

### Discussion

The StyleGAN model was trained on multimodal images containing CT and MR images of the pelvic region from male and female patients, with different pathologies. The FID score was about 12.3 which is similar to that for synthetically generated faces that Karras *et al.* achieved (they got 4.4). The data in this study differ from Karras *et al.* as the network had to learn two separate image distributions (*i.e.* MR and CT) that are very different in intensity and texture, and images at the boundary between the two modalities did not resemble images from either modality; these are likely two of the main factors contributing to a higher FID score. Further, comparing this metric for images of two different domains is challenging and has to be considered with caution. Nevertheless, the network

generates images that are of a high visual quality, and of a high resolution ( $512 \times 512$  pixels). Given the high visual quality of the images, it is likely that they can be used for training deep learning models, *e.g.* either for pre-training or as a form of data augmentation.

Affine transformations of latent style vectors appear to work well, and during this project we never found any "pockets" in the latent space where the model failed to generate realistic images. The StyleGAN model allows manipulation of individual features in the latent style space, W, and in particular it appears that the modality transfer, moving from MR to CT or the other way around, works satisfactory in practice. In fact, the reported MAE for transferring from the MR to the CT domain (73.6 HU) is not far from that reported in the literature on synthetic CT (sCT) generation, where the errors from using deep convolutional neural networks usually are in the range 40–50 HU [29–35].

The StyleGAN model further learned meaningful latent style space representations of the longitudinal direction of the patients. The errors grew with the distance from the initial slice, but were less than 75 HU MAE and 0.3 RMSE within 4 cm for CT and MR, respectively, for the range of factors that were tested. These are fairly small errors, as seen when comparing to the modality transfer, for instance.

The study was inconclusive about whether some principal directions have a strong curvature on the decision boundary, or if what we obtained is a result from biases in the reverse models. We can conclude that the hyper-surfaces of the dense neural networks appear to be mostly without curvature, which would explain why the logistic regression models performed almost as well as the dense neural networks. Our results do point towards the conclusion that the latent style space of the StyleGAN model is mostly disentangled, in the feature dimensions we have investigated, even if we cannot rule out entanglements.

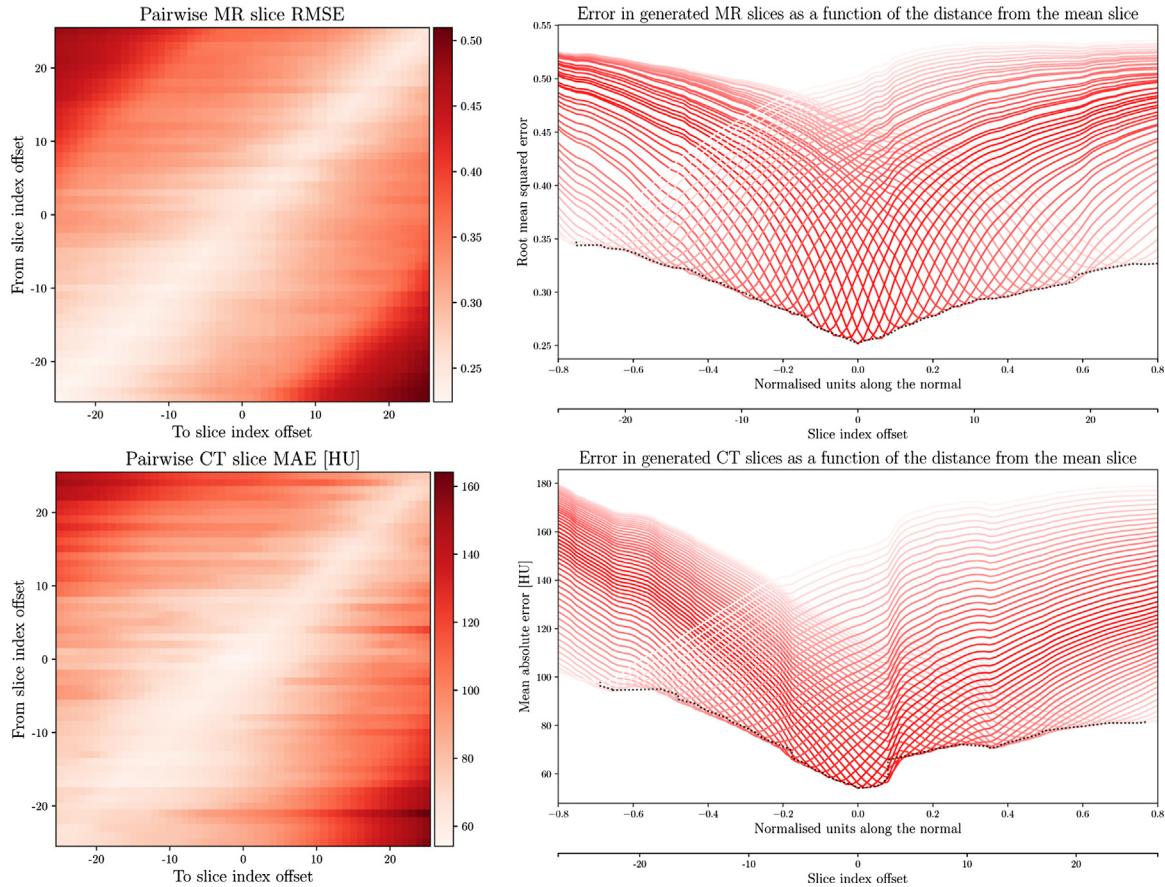


Figure 4. An illustration of moving from inferior to superior slices in the patient volume images. The top row corresponds to generated MR images, and the bottom row corresponds to generated CT images. The left plots illustrate the errors (the RMSE and MAEs over the 100 patient volumes, respectively) as we move from one slice (indices relative to the centre slice) to another slice (also with indices relative to the centre slice). The diagonal “valleys” imply that moving to nearby slices results in smaller errors compared to moving to distant slices. The right side of the plot (the line plots) illustrates the centre row from the corresponding left side of the plot (in the 2D plots). Each line illustrates the errors obtained as we move along the normal vector by a distance on the first axis. For each line, the errors are minimal when the sought slice is reached. Hence, the curve induced by the minima (highlighted with the black dotted line) of all the lines corresponds to the errors achieved as we move from the centre slice to farther away slices. The highlighted minima corresponds to the central row of the 2D plots. We note that the errors increase with the distance of the sought slice relative to the central slice.

Using the StyleGAN for style transfer appears to be another viable option for manipulating features. However, strong style transfers such as those demonstrated in the original paper were not observed during this study. This can be explained by the data arrangement of the original StyleGAN model where the output was three colour channels instead of one channel used in our work. Considering this architecture difference injecting latent vectors in the last layer cannot change the colour representation. The modality transfer had the smallest errors, and gave the qualitatively best results when the target latent vector was injected into the  $8 \times 8$  through  $64 \times 64$  resolution layers. The least MAE when generating CT images (58.6 HU) is close to that reported in the literature for sCT generation, and lower than the corresponding value from latent space manipulation. On the other hand, these generated images are likely

biased towards the paired image since those were used in this evaluation. Style mixing, where features of both style vectors (MR and CT) are strongly mixed, was only observed in the first resolution layer where both modality and longitudinal slice position changed simultaneously.

## Conclusion

The feature manipulation and style transfer capabilities of the StyleGAN makes it an attractive model to study the latent style space. The model that was presented in this study could be used to generate realistic slices, and possibly even volumes, of MR or CT images from synthetic patients of both sexes for training other deep learning models. The feasibility and benefits/drawbacks of using synthetic images such as those

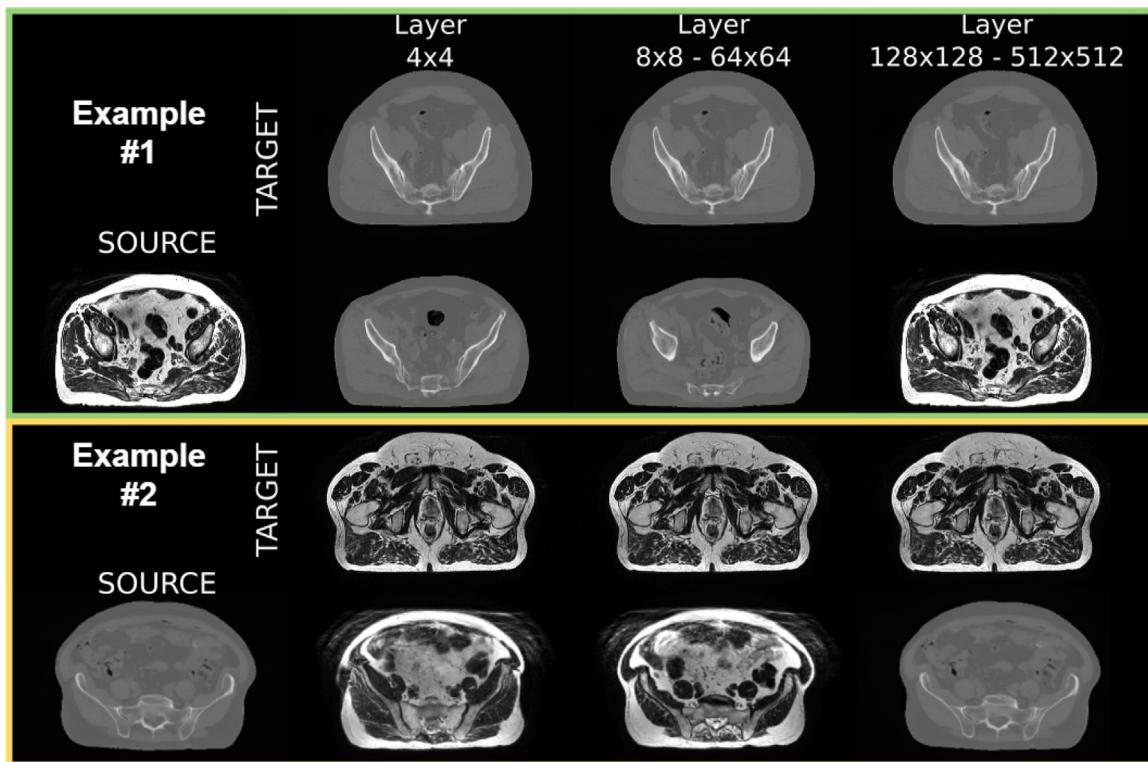


Figure 5. An illustration of style transfer with the StyleGAN, where features from MR (or respectively CT) were translated to CT (or MR, correspondingly), where the injection of the target style vector was in different locations of the StyleGAN (layers  $4 \times 4$ ,  $8 \times 8 - 64 \times 64$ , and  $128 \times 128 - 512 \times 512$ ) which are encoded in the columns. The green and orange boxes include two different examples where the green box includes the example of translating MR to CT and the orange box includes the example of translating CT to MR.

generated in this work for this purpose will be investigated in our future work.

Future work could also include automatic identification of feature directions, and means to further disentangle the latent style space for instance by regularising it. Such work and improvements could lead to better means to manipulate the images, and to generated images with entirely custom features. Further, the StyleGAN could be used to similarly investigate other body regions, such as the head for instance.

## Acknowledgement

Tufve Nyholm and Tommy Löfstedt are co-owners of NONPI Medical AB, the developer of MICE Toolkit—the software used in this work to prepare the training data.

This research was in part funded by the Austrian Science Fund (FWF, project number P30065-B27). Some of the GPUs used in this research were funded by a grant from the Cancer Research Fund of Northern Sweden. We gratefully acknowledge the support of Nvidia Corporation in their donation of a Titan Xp GPU used in this research.

## Appendix A Zusätzliche Daten

Zusätzliche Daten verbunden mit diesem Artikel finden sich in der Online-Version unter: <https://doi.org/10.1016/j.zemedi.2020.05.001>.

## References

- [1] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 2018;15(141).
- [2] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik* 2019;29(2):102–27.
- [3] Maier A, Syben C, Lasser T, Riess C. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik* 2019;29(2):86–101.
- [4] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. *Med. Phys* 2019;46(1):e1–36.
- [5] Feng Z, Nie D, Wang L, Shen D. Semi-supervised learning for pelvic mr image segmentation based on multi-task residual fully convolutional networks. *ISBI* 2018:885–8.
- [6] Jacobsen N, Deistung A, Timmann D, Goericke SL, Reichenbach JR, Güllmar D. Analysis of intensity normalization for optimal

- segmentation performance of a fully convolutional neural network. *Zeitschrift für Medizinische Physik* 2019;29(2):128–38.
- [7] Mahapatra D, Bozorgtabar B, Garnavi R. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics* 2019;71:30–9.
  - [8] Leynes AP, Yang J, Wiesinger F, Kaushik SS, Shanbhag DD, Seo Y, et al. Direct PseudoCT Generation for Pelvis PET/MRI Attenuation Correction using Deep Convolutional Neural Networks with Multi-parametric MRI: Zero Echo-time and Dixon Deep pseudoCT (ZeDD- CT). *J. Nucl. Med* 2017.
  - [9] Schnurr AK, Chung K, Russ T, Schad LR, Zöllner FG. Simulation-based deep artifact correction with Convolutional Neural Networks for limited angle artifacts. *Zeitschrift für Medizinische Physik* 2019;29(2):150–61.
  - [10] Russ T, Goerttler S, Schnurr AK, Bauer DF, Hatamikia S, Schad LR, Zöllner FG, Chung K. Synthesis of CT images from digital body phantoms using CycleGAN. *International Journal of Computer Assisted Radiology and Surgery* 2019;14(10):1741–50.
  - [11] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. *NIPS* 2014:2672–80.
  - [12] Burlina PM, Joshi N, Pacheco KD, Liu TY, Bressler NM. Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration. *JAMA Ophthalmology* 2019;137(3):258–64.
  - [13] Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. *ISBI* 2018.
  - [14] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for Medical Image Analysis. Preprint ariv: 2018, 1809.06222v2.
  - [15] Diaz-Pinto A, Colomer A, Naranjo V, Morales S, Xu Y, Frangi AF. Retinal Image Synthesis and Semi-supervised Learning for Glaucoma Assessment. *IEEE Transactions on Medical Imaging* 2019;38(9):2211–8.
  - [16] 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;321:321–31.
  - [17] Karras T, Aila T, Laine S, Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. in: *ICLR* 2018.
  - [18] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. *CVPR* 2019.
  - [19] Shen Y, Gu J, Tang X, Zhou B. Interpreting the Latent Space of GANs for Semantic Face Editing. Preprint ariv: 2019, 1907.10786.
  - [20] Abdal R, Qin Y, Wonka P. Image 2S tyleGAN: How to Embed Images Into the StyleGAN Latent Space? Preprint ariv: 2019, 1904.03189.
  - [21] Huang X, Belongie S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. *ICCV* 2017.
  - [22] Paszke A, Chanan G, Lin Z, Gross S, Yang E, Antiga L, et al. Automatic differentiation in PyTorch. *NIPS* 2017.
  - [23] Kingma DP, Ba JL. Adam: A method for stochastic optimization. in: *ICLR* 2017.
  - [24] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved Training of Wasserstein GANs. *NIPS* 2017.
  - [25] Mescheder L, Geiger A, Nowozin S. Which Training Methods for GANs do actually Converge? *ICML* 2018.
  - [26] 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. *NIPS* 2017.
  - [27] Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. *CVPR* 2017.
  - [28] Liu L, Jiang H, He P, Chen W, Liu X, Gao J, et al. On the Variance of the Adaptive Learning Rate and Beyond. Preprint ariv: 2019, 1908.03265.
  - [29] Korhonen J, Kapanen M, Keyriläinen J, Sepälä T, Tenhunen M. A dual model HU conversion from MRI intensity values within and outside of bone segment for MRI-based radiotherapy treatment planning of prostate cancer. *Medical Physics* 2013;41(1):011704.
  - [30] Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiation Oncology* 2017;12(1):28.
  - [31] Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA, İsgum I. Deep MR to CT synthesis using unpaired data. *MICCAI* 2017;14–23.
  - [32] Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks. *Deep Learn Data Label Med Appl* 2016;2016:170–8.
  - [33] Maspero M, Savenije MHF, Dinkla AM, Seevinck PR, Intven MPW, Jürgenliemk-Schulz IM, et al. Fast synthetic CT generation with deep learning for general pelvis MR-only Radiotherapy. *Phys Med Biol* 2018;1–14.
  - [34] Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst C. Generating Synthetic CTs from Magnetic Resonance Images using Generative Adversarial Networks. *Med Phys* 2018.
  - [35] Xiang L, Wang Q, Nie D, Zhang L, Jin X, Qiao Y, Shen D. Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. *Med Image Anal* 2018;47,: 31–44.

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

**ScienceDirect**