3D (c)GAN for Whole Body MR Synthesis

Daniel Mensing[1,2(✉)], Jochen Hirsch[1], Markus Wenzel[1], and Matthias Günther[1,2,3]

[1] Fraunhofer MEVIS, Bremen, 28359 Bremen, Germany
daniel.mensing@mevis.fraunhofer.de
[2] mediri GmbH, 69115 Heidelberg, Germany
[3] Universität Bremen, 28359 Bremen, Germany

Abstract. Synthesis of images has recently seen many works that produce high-quality real world images. In the domain of medical imaging the application of deep generative models especially Generative Adversarial Networks (GANs) can be applied to many different tasks. Under the premise of the generation of high-quality images that match the distribution of the original data, the synthesized data can be used to increase the size of small datasets, or in combination with conditioning on meta data, to increase the size of underrepresented classes in the dataset. In this work we propose a model that generates 3D medical images. The model can easily be conditioned on meta data, for example available patient information. We evaluate the quality of the generated images and compare our model against the 3D-StyleGAN model which is also designed for 3D medical image synthesis.

Keywords: Generative adversarial networks · 3D Image Synthesis · Conditional GAN

1 Introduction

In this work we propose a GAN architecture for the generation of 3D volumetric images. The design decisions of the architecture were inspired by the findings of DCGAN [16] and FastGAN [14] which were then validated for 3D medical image synthesis through an ablation study. Additionally we propose to use linear conditioning in the generator and discriminator on available meta data. There is little work on 3D medical image synthesis, especially with high resolution greater than 64^3. This can partly be explained with the requirement of Graphical Processing Unit (GPU) memory imposed by the three dimensions of the data. Often this lack of GPU memory has to be compensated by reducing the number of feature maps or the depth of the network which makes the training more challenging. Some previous works tried to overcome this issue by synthesizing only a slab of the volume [5] or generating the slices of the volume separately [2]. Previous work that generate volumes directly by using 3D convolutions is often limited in

A. Mukhopadhyay et al. (Eds.): DGM4MICCAI 2022, LNCS 13609, pp. 97–105, 2022.
https://doi.org/10.1007/978-3-031-18576-2_10

size/resolution. The authors of 3D-StyleGAN build upon the well-known StyleGAN2 architecture and adapted it for three dimensions by significantly reducing the number of feature maps and the size of the latent vector, to generate T1 weighted MR images at 2 mm spatial resolution [7]. We investigated previously known best practices for GANs and evaluated their feasibility for 3D medical images through an ablation study. As a result this work proposes a GAN that generates synthetic whole body MR volumes with a size of 160 × 160 × 128. We achieve this by reducing each training batch to a single data sample which allows us to increase the number of feature maps. Additionally we show that the proposed model can easily be conditioned on meta data which further improves its performance. We compare our model, with and without conditioning, with the 3D-StyleGAN architecture.

2 Methods

Architecture and Training

We propose an architecture based on findings for effective GAN training and adapted them to and investigated them for the domain of 3D medical images. The overall architecture and many parts were introduced by FastGAN [14]. Figure 1 shows a simplified diagram of the architecture. One major motivation behind this decision is the low demand for training data by the FastGAN in combination with the low complexity of the network. All design choices were validated through an ablation study in which we investigated the influence of each part on the models performance. We used InstanceNorm [17] instead of batch normalization because the size of the data does not allow for large batch sizes thus rendering batch normalization less useful. The generator first maps the latent vector to the first feature maps which determine the size of the output through a transposed convolution layer. The main building block of the generator is depicted in Fig. 2 on the left. The remaining generator consists of five of these blocks, each of which doubles the resolution of the intermediate feature maps and a final convolutional layer which maps the feature maps to the number of output channels, in this case one channel for grayscale images.

The discriminator mirrors this architecture except that the resolution of the feature maps is reduced by a factor of two by using strided convolutions and that the activation function for each convolution is the Leaky ReLU function. Furthermore, there is no noise injection in the discriminator. The repeating building blocks of the discriminator are shown in Fig. 2 on the right. At the end of the discriminator the features are fed into a small convolutional network consisting of two layers which reduces the size further and serves as a critic whose output rates the input data as real or fake which then is used for the adversarial training.

Both, generator and discriminator employ Skip-Layer-Excitation layers, introduced in [14], which serve as a skip connection between two blocks at different depth of the network and helps to propagate the error to the first layers of the model. Another important part of the decoder is self-regularization due to

decoders that decode the volume from the smallest feature map back to a volume with half the input size. This method was also introduced in [14] but we employ multiple decoders. We implemented one decoder, that decodes the feature maps to the whole volume, one that decodes to only one part of the input volume and one that only decodes the abdomen section to ensure high detail in this region. The decoder networks use transposed convolutions for the upsampling and no conditioning on meta data regardless of the conditioning in the generator and discriminator. The loss for the generator is the output of the critic, while the loss for the discriminator is the sum of the adversarial hinge loss [13] and separate reconstruction losses for each decoder, which were the mean absolute error between the decoded image and the interpolated or cropped part of the real image. Other methods used during the training process were exponential moving average of the generator weights [19], early stopping and learning rate decay.

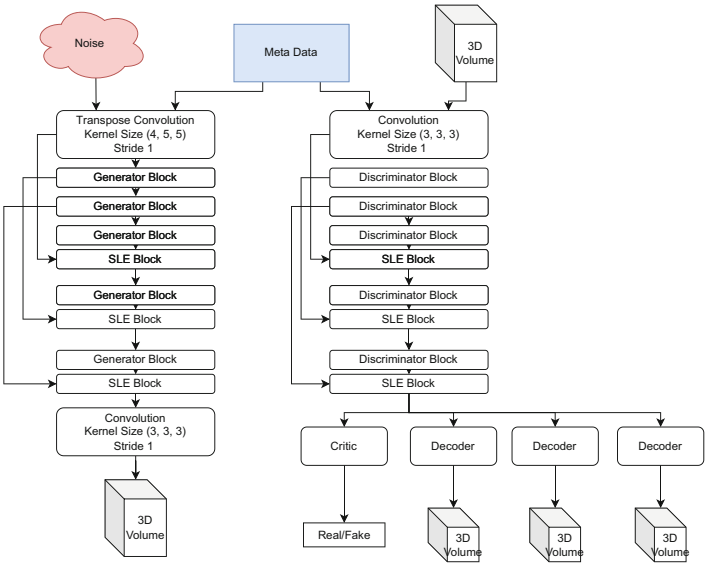

Fig. 1. Architecture of the Generator (left) and Discriminator (right) networks. Simplified view of the architecture of our proposed GAN. The meta data input for both models is optional. A detailed view of the generator and discriminator blocks is provided in Fig. 2. In general the architecture is inspired by the overall architecture proposed in [14]. Skip Layer Excitation (SLE) blocks are used to propagate the error to the first layers of the model.

Conditioning on Meta data

Many use cases benefit from the ability to generate data conditioned on given attributes. The following patient information were used for the conditional 3D

image synthesis: *age*, *sex*, *weight* and *height*. For conditioning, we added a Feature-wise Linear Modulation (FiLM) layer [15] between each convolutional layer and the noise injection layer. This layer affine transforms the intermediate feature maps with two learned parameter vectors γ and β, which are provided by an encoder, which is trained together with the model, that is shared through all FiLM layers in the model (generator and discriminator each have their own). For this experiment, we binary-encoded the meta data and concatenated all binary vectors which then serves as input for the encoder. If the network shall be conditioned on additional input data, a FiLM modulation layer follows between the convolution layer and the noise injection layer. The linear conditioning with meta data was shown to be beneficial for image segmentation by [12].

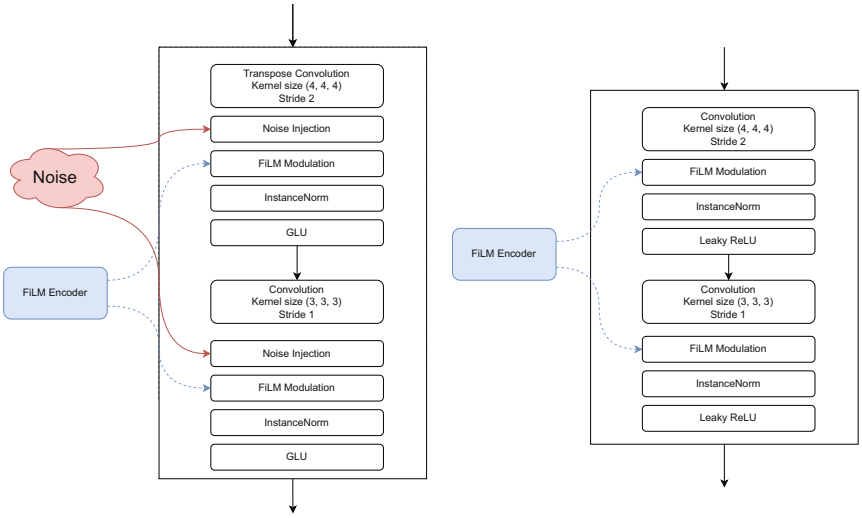

Fig. 2. Architecture of the Generator (left) and Discriminator (right) blocks. Each block doubles the size if the incoming feature maps in each dimension through a transposed convolution, then adds a random sampled noise vector of the same size as the feature maps which has been proven beneficial for the circumvention of overfitting and improving the generalization [4]. The resulting output is normalized by an Instance Normalization [17] layer and a Gated Linear Unit (GLU) [3] operation serves as activation function. The same structure is repeated once more with the transposed convolution replaced by a regular convolution

Data

The used dataset consists of 10828 whole body MR volumes obtained as part of the MR Imaging Study within the German National Cohort Study (GNC, 2014-2019) [1] from volunteers. The data was acquired on MAGNETOM Skyra 3T (Siemens Healthineers, syngo VD13C) systems with a two-point Dixon volumetric interpolated breath-hold examination (VIBE) T1 weighted sequence. We

used the so-called "opposed phase" contrast (TE = 1.23 ms). The volumes were acquired by axial acquisition with in-plane matrix 320 × 260 (resolution 1.4 × 1.4 mm^2) and a slice thickness of 3 mm. The volume consists of four acquired table positions with a total of 316 slices which were then resampled and cropped to 160 × 160 × 128 which doubles the voxel size but therefore reduces the size of each volume roughly by half in order to fit the volume on the GPU. All intensity values were scaled to the range of [–1, 1]. We used half of the dataset for training and the other half for evaluation.

Evaluation

For the evaluation of the quality of the generated volumes we used the slice-wise Fréchet Inception Metric proposed in [7]. Since the Fréchet Inception Distance (FID) [6] is calculated from features extracted from a Inception V3 network pretrained on the Imagenet dataset, which is a 2D dataset, we calculate the FID score for the center slice for each orientation. Additionally, the Multi-Scale Structured Similarity Measure (MS-SSIM) and the Maximum Mean Discrepancy (MMD) were used for the evaluation. The MMD measures the distance between two distributions and was calculated batch-wise as proposed in [11,18] and [7]. The MS-SSIM measures the structural similarity between two samples at different scales and can be used to evaluate the diversity of the generated images [18].

3 Results

The results shown in Table 1 show that our model without conditioning has a much lower MMD and FID and higher MS-SSIM than the trained 3D-StyleGAN. A comparison between a sample generated by the 3D-StyleGAN and our unconditional model is shown in Fig. 3. In comparison our model with conditioning on meta data results in even slightly better scores across almost all metrics. The 3D-StyleGAN was trained with 1 mm-fd16 configuration which was the only one that allows to generate volumes at the size of 160 × 160 × 128. The only change to the configuration was the output size of the base layer which was adapted from 5 × 6 × 7 to 5 × 5 × 4 to result in the desired output size.

Table 1. Results. The table shows the MMD, the MS-SSIM between whole volumes of generated and real data. The FID was calculated for the center slice of the volume in Axial (FID Ax.), Sagittal (FID Sag.) and Coronal (FID Cor.) orientation between generated and real samples. ↓ means that a lower metric score is better and ↑ shows that a higher value is better.

	MMD ↓	MS-SSIM ↑	FID (Ax.) ↓	FID (Sag.) ↓	FID (Cor.) ↓
3D-StyleGAN	47307 ± 13162	0.162 ± 0.004	362.5 ± 1.6	373.7 ± 15.9	431.6 ± 11
Ours	12086 ± 641	0.409 ± 0.004	**71.2** ± 1.0	43.3 ± 5.7	106.4 ± 23.7
Ours Conditional	**10589** ± 333	**0.439** ± 0.001	76.5 ± 2.5	**38.4** ± 10.2	**81.6** ± 22.5

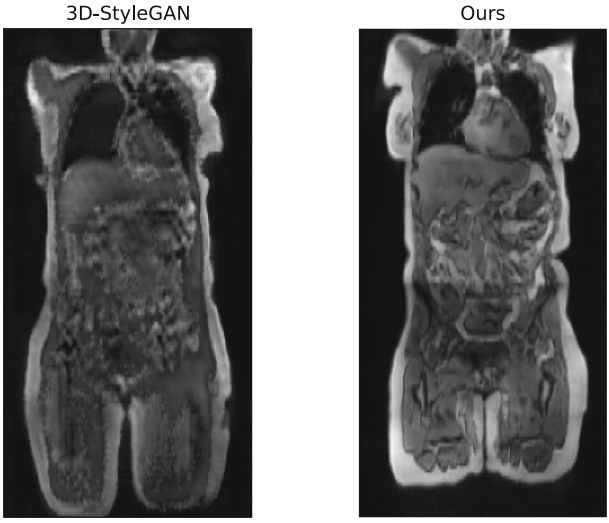

Fig. 3. Comparison between 3D-StyleGAN and ours. This figure shows the center slice in coronal orientation of two samples, on the left side generated by 3D-StyleGAN and on the right side by our proposed model. Both samples were generated unconditionally.

Conditional generation of 3D images

The results of the conditioning process were not evaluated separately. A visual inspection of the conditionally generated volumes showed that these were consistent with the meta data they were conditioned on which can be seen in Fig. 4. Depicted are generated volumes from the same latent vector and with different meta data conditioning. The images show the center slice of male and female volumes with different weights from 60 to 110 Kg. The other two remaining attributes stayed fixed.

4 Discussion

We propose a GAN architecture for 3D medical image synthesis that uses best practices for GAN training known from other domains. In order to leverage the often limited datasets available for medical imaging we added self regularization by adding decoders to the discriminator as proposed and justified in [14]. We assess our models performance with commonly used metrics for the evaluation of GANs and compare these against the 3D-StyleGAN architecture at the same resolution. In Table 1 we show that our model outperforms the 3D-StyleGAN in every metric. A possible explanation of the in general low MS-SSIM score across all compared models may be partially explained by the fact, that the training data has not been registered and therefore exhibits variation in size and scale of the samples. Since the MS-SSIM compares the structural similarity of spatially

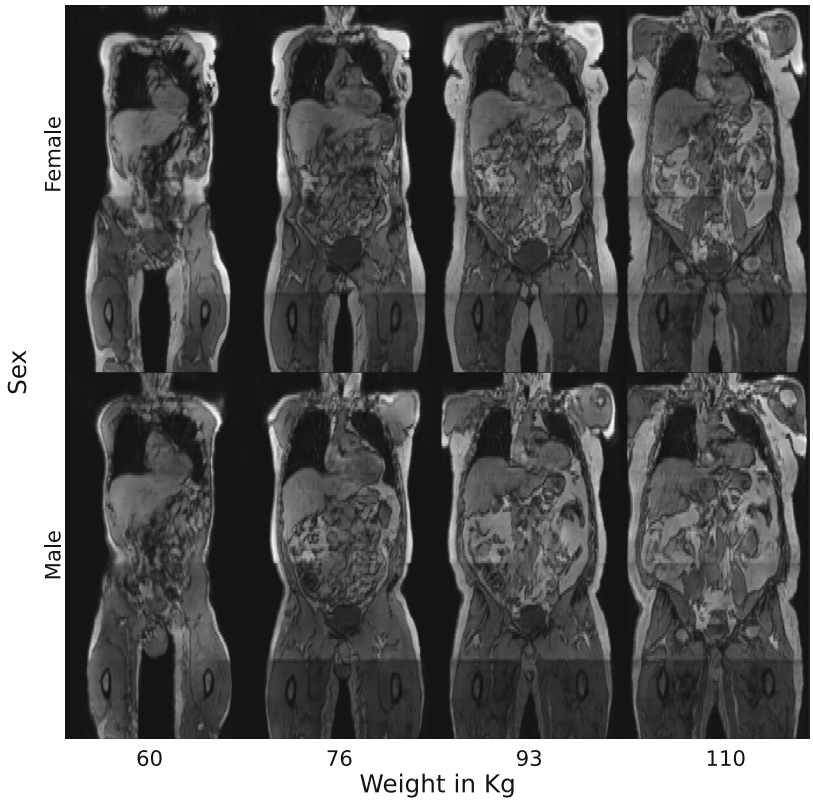

Fig. 4. Conditionally generated volumes. Center slices of volumes generated from the same latent vector with different conditional inputs. The upper row shows samples for the attribute *sex* set to female and the lower row with the attribute set to male. Both rows show the variation for the *weight* attribute ranging from 60 to 110 Kg. The remaining attributes *height* and *age* were set to 170 cm and 40 years old respectively. The stitching artefact that can be seen was caused by movements between the four acquisitions of which the volume is put together and was also learned by the GAN.

close voxels and the corresponding voxel in two compared samples can be at spatially different locations due to the patients size, this is a possible explanation for low MS-SSIM scores. [7] argued their model was not able to generate realistic images at 1 mm isotropic resolution which translates to an image size of $160 \times 192 \times 224$. Since the size of our images ($160 \times 160 \times 128$) is in between the size of their successful experiments and their failure case, we can only deduct that our model results in lower metric scores at the reported size. The conditioning on patient information has to be investigated further in regard of the independence of the different attributes and if the conditionally generated 3D images are plausible for the used meta data. Further improvements to the proposed architecture could stem from the StyleGAN [8–10] models which propose differentiable data

augmentation or weight modulation. Further experiments with other modalities, organs and image sizes are needed to show the ability of the model to generalize beyond the trained data. Very recently HA-GAN was published in which the authors synthesize chest CT and brain MR images with a size of 256^3 and a comparison is left for future work.

Acknowledgement. We received grant money from the U Bremen Research Alliance/AI Center for Health Care, financially supported by the Federal State of Bremen.

References

1. Bamberg, F., et al.: Whole-body MR imaging in the German national cohort: rationale, design, and technical background. Radiology **277**(1), 206–220 (2015)
2. Bergen, R.V., Rajotte, J.F., Yousefirizi, F., Klyuzhin, I.S., Rahmim, A., Ng, R.T.: 3D PET image generation with tumour masks using TGAN. In: Medical Imaging 2022: Image Processing, vol. 12032, p. 120321P (2022). https://doi.org/10.1117/12.2611292
3. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. arXiv (2016)
4. Feng, R., Zhao, D., Zha, Z.: On noise injection in generative adversarial networks. arXiv (2020)
5. Granstedt, J.L., Kelkar, V.A., Zhou, W., Anastasio, M.A.: SlabGAN: a method for generating efficient 3D anisotropic medical volumes using generative adversarial networks. In: Medical Imaging 2021: Image Processing, vol. 11596, p. 1159617 (2021). https://doi.org/10.1117/12.2581380
6. 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. arXiv (2017)
7. Hong, S., et al.: 3D-StyleGAN: a style-based generative adversarial network for generative modeling of three-dimensional medical images. arXiv (2021)
8. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv (2020)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. arXiv (2018)
10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. arXiv (2019)
11. Kwon, G., Han, C., Kim, D.S.: Generation of 3D brain MRI using auto-encoding generative adversarial networks. arXiv (2019)
12. Lemay, A., Gros, C., Vincent, O., Liu, Y., Cohen, J.P., Cohen-Adad, J.: Benefits of linear conditioning with metadata for image segmentation. arXiv (2021)
13. Lim, J.H., Ye, J.C.: Geometric GAN (2017). https://doi.org/10.48550/ARXIV.1705.02894, https://arxiv.org/abs/1705.02894
14. Liu, B., Zhu, Y., Song, K., Elgammal, A.: Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis. arXiv (2021)
15. Perez, E., Strub, F., Vries, H.D., Dumoulin, V., Courville, A.: FiLM: visual reasoning with a general conditioning layer. arXiv (2017)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv (2015)

17. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. CoRR abs/1607.08022 (2016), http://arxiv.org/abs/1607.08022
18. Volokitin, A., et al.: Modelling the distribution of 3D brain MRI using a 2D slice VAE. arXiv (2020)
19. Yazıcı, Y., Foo, C.S., Winkler, S., Yap, K.H., Piliouras, G., Chandrasekhar, V.: The unusual effectiveness of averaging in GAN training. arXiv (2018)