Medical Image Analysis 78 (2022) 102396

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Generating 3D TOF-MRA volumes and segmentation labels using generative adversarial networks

Pooja Subramaniam[a], Tabea Kossen[a,b,*], Kerstin Ritter[c,d], Anja Hennemuth[b,e,f], Kristian Hildebrand[g], Adam Hilbert[a], Jan Sobesky[h,i], Michelle Livne[a], Ivana Galinovic[i], Ahmed A. Khalil[i,j,k,l], Jochen B. Fiebach[i], Dietmar Frey[a], Vince I. Madai[a,m,n]

[a] CLAIM - Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Germany
[b] Department of Computer Engineering and Microelectronics, Computer Vision & Remote Sensing, Technical University Berlin, Berlin, Germany
[c] Department of Psychiatry and Psychotherapy, Charité Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Berlin, Germany
[d] Bernstein Center for Computational Neuroscience, Berlin, Germany
[e] Institute for Imaging Science and Computational Modelling in Cardiovascular Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany
[f] Fraunhofer MEVIS, Max-von-Laue-Str. 2, Bremen, Germany
[g] Department VI Computer Science and Media, Beuth University of Applied Sciences, Berlin, Germany
[h] Johanna-Etienne-Hospital, Neuss, Germany
[i] Centre for Stroke Research Berlin, Charité Universitätsmedizin Berlin, Berlin, Germany
[j] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[k] Mind, Brain, Body Institute, Berlin School of Mind and Brain, Humboldt University Berlin, Berlin, Germany
[l] Berlin Institute of Health, Berlin, Germany
[m] School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, UK
[n] QUEST-Center for Transforming Biomedical Research, Berlin Institute of Health, Charité Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany

ARTICLE INFO

Article history:
Received 13 July 2021
Revised 28 January 2022
Accepted 17 February 2022
Available online 24 February 2022

MSC:
41A05
41A10
65D05
65D17

Keywords:
Generative adversarial networks
3D Medical imaging
Mixed precision
Anonymization
Brain vessel segmentation

ABSTRACT

Deep learning requires large labeled datasets that are difficult to gather in medical imaging due to data privacy issues and time-consuming manual labeling. Generative Adversarial Networks (GANs) can alleviate these challenges enabling synthesis of shareable data. While 2D GANs have been used to generate 2D images with their corresponding labels, they cannot capture the volumetric information of 3D medical imaging. 3D GANs are more suitable for this and have been used to generate 3D volumes but not their corresponding labels. One reason might be that synthesizing 3D volumes is challenging owing to computational limitations. In this work, we present 3D GANs for the generation of 3D medical image volumes with corresponding labels applying mixed precision to alleviate computational constraints.

We generated 3D Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) patches with their corresponding brain blood vessel segmentation labels. We used four variants of 3D Wasserstein GAN (WGAN) with: 1) gradient penalty (GP), 2) GP with spectral normalization (SN), 3) SN with mixed precision (SN-MP), and 4) SN-MP with double filters per layer (c-SN-MP). The generated patches were quantitatively evaluated using the Fréchet Inception Distance (FID) and Precision and Recall of Distributions (PRD). Further, 3D U-Nets were trained with patch-label pairs from different WGAN models and their performance was compared to the performance of a benchmark U-Net trained on real data. The segmentation performance of all U-Net models was assessed using Dice Similarity Coefficient (DSC) and balanced Average Hausdorff Distance (bAVD) for a) all vessels, and b) intracranial vessels only.

Our results show that patches generated with WGAN models using mixed precision (SN-MP and c-SN-MP) yielded the lowest FID scores and the best PRD curves. Among the 3D U-Nets trained with synthetic patch-label pairs, c-SN-MP pairs achieved the highest DSC (0.841) and lowest bAVD (0.508) compared to the benchmark U-Net trained on real data (DSC 0.901; bAVD 0.294) for intracranial vessels.

In conclusion, our solution generates realistic 3D TOF-MRA patches and labels for brain vessel segmentation. We demonstrate the benefit of using mixed precision for computational efficiency resulting in the best-performing GAN-architecture. Our work paves the way towards sharing of labeled 3D medical data which would increase generalizability of deep learning models for clinical use.

https://doi.org/10.1016/j.media.2022.102396

1. Introduction

The success of deep learning algorithms in natural image analysis has been leveraged in recent years to the medical imaging domain. Deep learning methods have been used for automation of various manual time-consuming tasks such as segmentation and classification of medical images (Greenspan et al., 2016; Lundervold and Lundervold, 2019). Supervised deep learning methods, specifically, learn relevant features from images by mapping features in the input images to the label output. While the advantage of these methods is that they do not need manual extraction of features from the images, they do require large amounts of labeled data. Here, a major challenge is that it is expensive and difficult to acquire and label medical data (Yi et al., 2019). Yet, even when labeled medical data is available, it usually cannot be shared readily with other researchers due to privacy concerns (Clinical Practice Committee, 2000). Anonymization methods typically applied in medical imaging would not be beneficial in the case of neuroimaging as the unique neuroanatomical features present in brain images could be used to identify individuals (Wachinger et al., 2015; Valizadeh et al., 2018). As a consequence, often small, siloed or homogenous datasets are used when proposing new deep learning models in neuroimaging (Willemink et al., 2020).

A potential solution to this problem is the generation of synthetic medical imaging data. A very promising method for this purpose is Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Various GAN architectures from the natural images domain have gained popularity in medical imaging for image synthesis, supervised image-to-image translation, reconstruction and super-resolution (Yi et al., 2019). For image synthesis, specifically, 2D GANs have been used in several works such as synthesis of Computed Tomography (CT) liver lesions (Frid-Adar et al., 2018), skin lesion images (Baur et al., 2018), and axial Magnetic Resonance (MR) slices (Bermudez et al., 2018). GANs can be extended to generate the labels along with the synthesized images. For example, 2D GANs have been used to generate the corresponding segmentation labels for lung X-rays (Neff et al., 2018), vessel segmentation (Kossen et al., 2021), retinal fundus images (Guibas et al., 2018) and brain tumor segmentation (Foroozandeh and Eklund, 2020). Although these results are promising, the challenge remains that 2D GANs cannot capture important anatomical relationships in the third dimension. Since medical images are often recorded in 3D, GANs generating 3D medical images are thus highly warranted. 3D GANs have been used to generate downsampled or resized MRI images of different resolutions (Kwon et al., 2019; Eklund, 2020; Sun et al., 2021). However, to our knowledge, there is no 3D GAN medical imaging study that generates the corresponding labels, which is critical for using the data for supervised deep learning research. One reason could be that synthesizing 3D volumes is still a challenge due to computational limitations.

In our study, we generate high resolution 3D medical image patches along with their labels in an end-to-end paradigm for brain vessel segmentation which aids in identifying and studying cerebrovascular diseases. From 3D Time-of-Flight Magnetic Resonance Angiography (TOF-MRA), we synthesize 3D patches together with brain vessel segmentation labels. We implement and compare four different 3D Wasserstein-GAN (WGAN) variants: three with the same architecture but different regularizations and mixed precision (Micikevicius et al., 2018) schemes, and one with a modified architecture - double filters per layer - owing to memory efficiency from mixed precision. Next to a qualitative visual assessment, we use quantitative measures to evaluate the synthesized patches. We further evaluate the performance of brain vessel segmentation models trained on the generated patch-label pairs and compare them to a benchmark model trained on real data. Additionally, we also compare the segmentation performance on a second, independent dataset.

To summarize, our main contributions are:

1. For the first time to our knowledge in the medical imaging domain, we generate high resolution 3D patches along with segmentation labels using GANs.
2. We utilize the memory efficiency provided by mixed precision to enable a more complex WGAN architecture with double the filters per layer.
3. Our generated labels allow us to train 3D U-Net models for brain vessel segmentation on synthetic data in an end-to-end framework.

2. Methods

2.1. Architecture

We adapted the WGAN - Gradient penalty (Gulrajani et al., 2017) model to 3D in order to produce 3D patches and their corresponding labels of brain vessel segmentation. We implemented four variants of the architecture: a) *GP model* - WGAN-GP model in 3D b) *SN model* - GP model with spectral normalization in the critic network c) *SN-MP model* - SN model with mixed precision d) *c-SN-MP model* - SN-MP model with double the filters per layer. An overview of the GAN training is provided in Fig. 1.

For all models, a noise vector (z) of length 128 sampled from a standard Gaussian distribution ($\mathcal{N}(0, 1)$) was input to the Generator G. It was fed through a linear layer and a 3D batch normalization layer, then 3 blocks of upsampling and 3D convolutional layers with consecutive batch normalization and ReLU activation, and a final upsampling and 3D convolutional layer as shown in Fig. 2A. An upsample factor of 2 with nearest neighbor interpolation was used. The convolutional layers used kernel size of 3 and stride of 1. Hyperbolic tangent ($tanh$) was used as the final activation function. The output of the generator was a two channel image of size $128 \times 128 \times 64$: one channel was the TOF-MRA patch and the second channel was the corresponding label which is the ground truth segmentation of the generated patch. The function of the labels is to train a supervised segmentation model such as a 3D U-Net model with the generated data.

Next, the critic D either took the generated 3D patch-label pairs ($G(z(i))$) or the real 3D patch-label pairs (x) as its input. The patch-label pairs were fed through four 3D convolutional layers. A kernel size of 3 and stride of 2 was used in the convolutional layers. After each convolutional layer, a 3D instance normalization layer was used for the GP model as shown in Fig. 2B. Here, for the SN model, we used spectral normalization (Miyato et al., 2018) after each convolutional layer which acts as an additional regularization to gradient penalty as shown in Fig. 2C. Leaky ReLU was used as the activation layer after the normalization layers. The last layer was linear that produced a scalar, coined as the critic's score. The score indicates how similar the distribution of the generated patch-label pairs is to that of the real patch-label pairs. This indirectly ensures that the generated labels correspond to the vessels in the generated patches similar to how the real labels correspond to the vessels in the real patches. The loss function of the critic was:

$$loss_D(i) = D(G(z^{(i)})) - D(x) + \lambda(\|\nabla D(\hat{x})\| - 1)^2 \quad (1)$$

where $\hat{x} = \epsilon x + (1 - \epsilon)G(z^{(i)})$, $\epsilon \sim U[0, 1]$, $\lambda = 10$ and ∇ is gradient of the critic. Here, the difference between the critic's score for the

* Corresponding author at: CLAIM - Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Germany.
E-mail address: tabea.kossen@charite.de (T. Kossen).

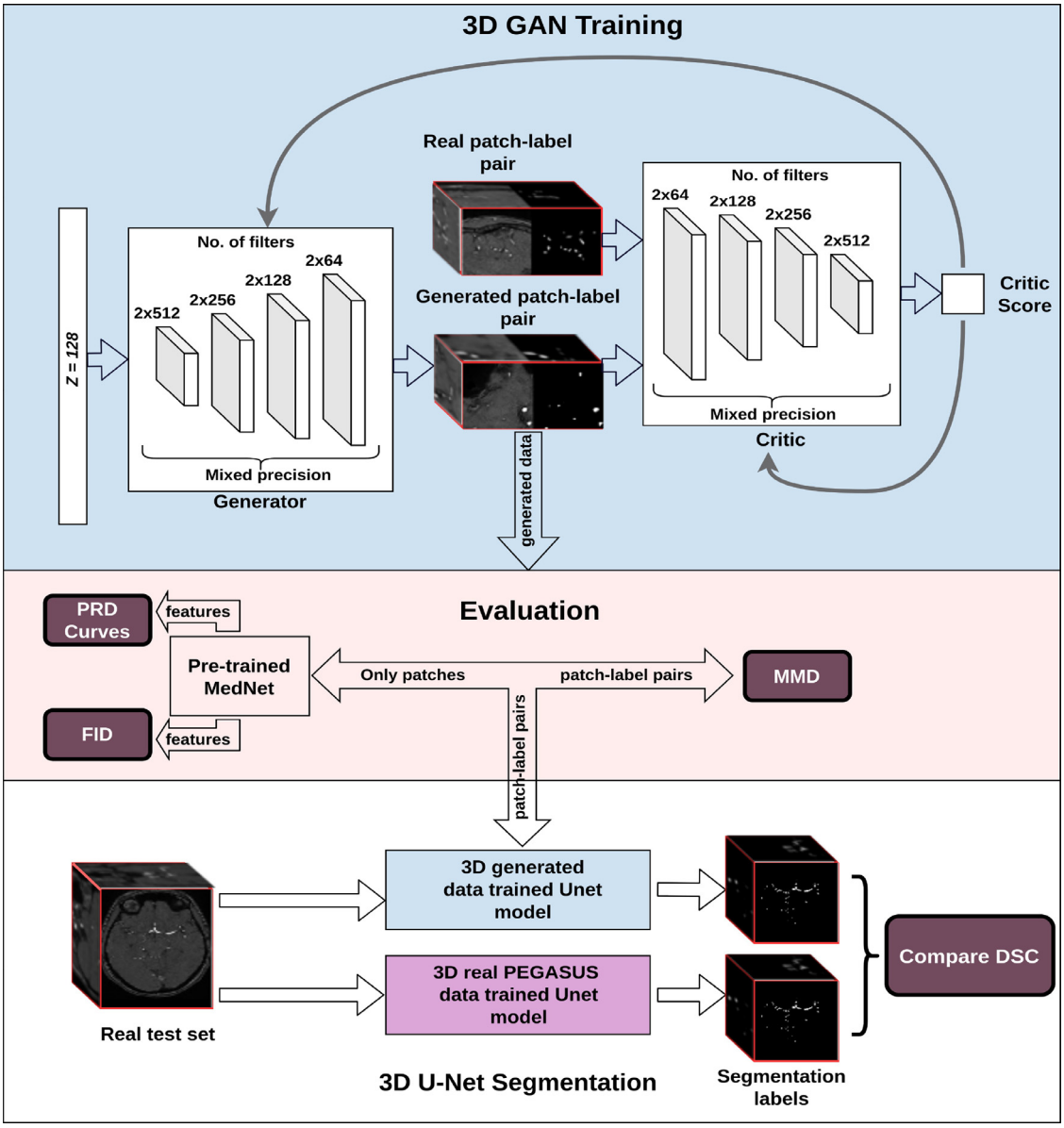

Fig. 1. Structure of the workflow from training the 3D GAN to qualitative and quantitative assessments. Top: Overview of GAN training - Here, we illustrate our most complex model using spectral normalization and mixed precision (c-SN-MP), middle: Evaluation schemes, bottom: Segmentation performance evaluation.

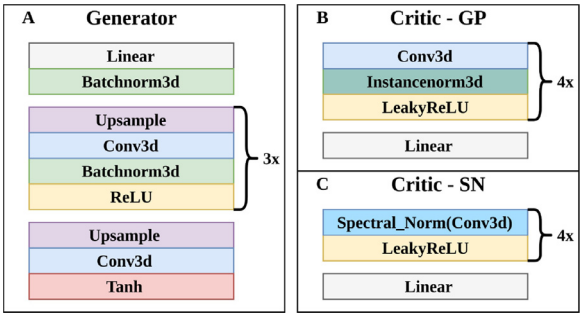

Fig. 2. Architectures of A. Generator of all models, B. Critic of GP model, and C. Critic of all SN models.

real and generated data along with the gradient penalty is computed. The loss function of the generator based on the output of the critic was:

$$loss_G(i) = -D(G(z^{(i)})) \quad (2)$$

This equates to maximizing the critic's score for the generated images by using the negative of the critic's score as loss for the generator.

In the case of the SN-MP model, mixed precision was used for memory efficiency. The default precision used in deep learning methods is 32 floating point (FP32). In mixed precision, both half precision (FP16) and FP32 are used depending on the precision requirements of a particular arithmetic operation. Here, FP16 is used for storing weights, activations and gradients while an FP32 master copy of weights is used for optimizer updates. A loss-scaling factor is applied in order to maintain the performance equivalent to a fully FP32 network. Using mixed precision, allowed us to use more filters per layer. Hence, c-SN-MP model was trained where double the filters were used in each layer of the SN-MP model. For implementation details, see open source code[1].

[1] https://github.com/prediction2020/3DGAN_synthesis_of_3D_TOF_MRA_with_segmentation_labels

2.2. Data

2.2.1. Datasets

TOF-MRA data of 137 patients with cerebrovascular disease from two earlier studies, PEGASUS ($n = 72$) and 1000Plus ($n = 65$), were used in this work. The 65 TOF-MRA data from the 1000Plus study were used for additional validation as a second, independent dataset. The details of the studies can be found in (Mutke et al., 2014) for PEGASUS and in (Hotter et al., 2009) for 1000Plus. The imaging was performed with the following parameters for both the studies: voxel size=$0.5 \times 0.5 \times 0.7$ mm3 ; matrix size=$312 \times 384 \times 127$; TR/TE=22 ms/3.86 ms; time of acquisition=3:50 min, flip angle=18 degrees.

The images were pre-segmented semi-manually with a standardized pipeline using a thresholded region growing algorithm in the case of PEGASUS dataset, and a 2D U-Net segmentation model (Livne et al., 2019) in the case of 1000Plus. Final ground truths were created following pre-defined manual correction steps first by junior and finally senior raters. Further details of the labeling methodology can be found in (Hilbert et al., 2020).

2.2.2. Data splitting and preprocessing

TOF-MRA images from each study were denoised, and non-uniformity correction was applied to improve image quality (Masoudi et al., 2021). For training the GANs, 47 of the patient data of the PEGASUS study were used. For the downstream task of segmentation, 12 were used as the validation set and 13 as the test set. The 1000Plus study dataset was solely used as an independent test set ($n = 65$) for evaluation of the trained segmentation model.

Due to computational limitations, 3D patches of size $128 \times 128 \times 64$ were extracted from the whole brain TOF-MRA scans of the PEGASUS training set. For training of GANs, 50 patches of images and their labels per patient were extracted - in part systematically (18) to cover all parts of the image and in part randomly (32) with the center voxel being a blood vessel in order to represent sufficient vessels. This amounted to a total of 2,350 patch-label pairs. In addition, 250 patches per patient were randomly extracted with center voxel as blood vessel for the downstream segmentation model from the PEGASUS training and validation set leading to 11,750 and 3,000 patch-label pairs respectively.

The image patches for the GAN training were normalized between -1 and +1. The corresponding labels were stacked on the image patch as a second channel for training the GANs.

2.3. Evaluation methods

An overview of the evaluation methods is shown in Fig. 1. The qualitative evaluation was done by visually assessing the images, labels and the 3D vessel structure using ITK-SNAP[2] as a first step. For a quantitative assessment, FID scores were computed from the extracted features using MedicalNet following precedence (Sun et al., 2021). This is a 3D ResNet model pretrained on 23 different medical datasets for segmentation (Chen et al., 2019). We chose this network instead of the commonly used Inception-v3 trained on ImageNet dataset (Szegedy et al., 2016) for calculating the FID scores to better match our 3D medical data.

While the FID measures the quality of the images, it does not account for mode collapse. Mode collapse happens when the generator learns to output a small set of good quality images to get a good critic's score and does not learn further any new variations present in the training data. In order to quantify both quality and variety of modes captured in the synthetic data, we used Precision and Recall for Distributions (PRD) (Sajjadi et al., 2018). Precision quantifies the quality of the image, and Recall amounts to the mode collapse. We also computed the Area Under the Curve (AUC) of the PRD curves to extract a single score for a simple quantification. Here again, we compared the extracted features of the generated and real patches from the pre-trained MedicalNet. It is important to note that both FID and PRD curves are based on the imaging patches alone and the labels are not taken into consideration for these performance measures.

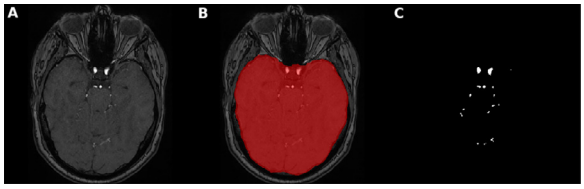

Fig. 3. Brain mask application for intracranial vessels analysis. Here, an axial slice is shown of A. TOF-MRA image with skull B. brain mask extracted using FSL-BET tool from TOF-MRA image C. ground truth segmentation label after brain mask application leading to skull-stripping i.e. removal of all vessels of face and neck with only intracranial vessels remaining.

Next, we tested the generated data for brain vessel segmentation. 3D U-Nets were trained on the synthetic patch-label pairs produced from the four different 3D GANs, and on the real data to compare segmentation performance. The generated patches were rescaled back to the real data range i.e. to 0–255 and the labels made binary by using a threshold. The performance of all trained U-Nets was evaluated on two independent test sets in two separate analysis schemes: a) all vessels b) intracranial vessels. In the case of all vessels, the whole predicted segmentation label was considered for evaluation. For intracranial vessels, the segmentation labels were processed so that only the intracranial vessels were considered. This was done by applying brain masks of corresponding TOF-MRA images on the ground truth segmentation labels and the prediction labels from all the U-Net models. The brain masks were obtained automatically using the FSL-Brain Extraction Tool[3] (BET) with parameter $frac = 0.05$ on the TOF-MRA images. A visual illustration of this post-processing of labels for intracranial vessels is shown in Fig. 3. In each case, the U-Net model that performed the best on the real validation set was selected to compute and report the performance on the real test sets. This method of evaluation not only signifies the utility of the synthetic data for the brain vessel segmentation use case but also provides information about how well the generated labels reflect the vessel information in the generated patch as this is crucial for a good segmentation performance. The segmentation performance was measured using Dice Similarity Coefficient (DSC) and the balanced Average Hausdorff Distance (bAVD) (Aydin et al., 2021b). DSC is a commonly used metric to evaluate segmentation performance, given by:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

where TP = True positive; FP = False positive; FN = False negative. A higher DSC indicates good segmentation performance. bAVD is a distance metric which has been shown to be a better metric for evaluation of blood vessel segmentation (Aydin et al., 2021a). It is a modified average Hausdorff distance defined as:

$$bAVD = \frac{1}{2} \times \left(\frac{1}{N_G} \sum_{g \in G} \min_{p \in P}(d(g,p)) + \frac{1}{N_G} \sum_{p \in P} \min_{g \in G}(d(p,g)) \right) \tag{4}$$

where G is the set of voxels in the ground truth, P is the set of voxels in the predicted segmentation. The balanced directed average Hausdorff distance from voxel set G to P is given by the sum of all minimum distances from all points belonging to point

[2] http://www.itksnap.org/pmwiki/pmwiki.php

[3] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET/UserGuide

set G to P divided by the number of points in G. Similarly, balanced directed average Hausdorff distance from voxel set P to G is given by the sum of all minimum distances from all points belonging to point set P to G divided by the number of points in G. bAVD is the mean of the directed average Hausdorff distance from G to P and directed average Hausdorff distance from P to G in voxels. A lower bAVD indicates good segmentation performance. We used the EvaluateSegmentation tool (Taha and Hanbury, 2015) to calculate the DSC and bAVD for each patient prediction. The mean DSC and mean bAVD was then calculated across all the patients.

2.4. Training

The models were implemented in PyTorch, and trained using an Nvidia TITAN RTX GPU for 100 epochs each. We used two time-scale update rule (Heusel et al., 2018) with different learning rates of 0.0004 and 0.0002 for the critic and the generator respectively instead of having more updates for the critic within each epoch. Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0$ and $\beta_2 = 0.9$ was used. The batch-size for all models was 4. For mixed precision, the Automatic Mixed Precision (AMP) package from PyTorch was used. A threshold of 0.3 was set for binarizing the generated labels except in the case of SN model where 0.2 was used. All the above hyperparameters were chosen based on the performance of the validation set in the segmentation task. The training times and the memory used for each GAN variant were recorded.

Table 1
FID scores and AUC of the PRD curves for synthetic data from different models.

Data source	FID	PRD-AUC
GP model	0.0381	0.80
SN model	0.0322	0.82
SN-MP model	**0.0206**	**0.87**
c-SN-MP model	0.0244	0.86

For segmentation, the published 3D U-Net architecture and framework implemented in TensorFlow from Hilbert et al. (2020) was utilized with the default hyperparameters. These were Adam optimizer with a learning rate of 0.0001 and $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size of 8.

3. Results

In the visual analysis, the synthetic patches, labels and the 3D vessel structure from the complex mixed precision model (c-SN-MP) appeared as the most realistic (Fig. 4). The patches from the mixed precision models (SN-MP and c-SN-MP) had the lowest FID scores (Table 1), and the best PRD curves (Fig. 5). Based on the PRD curves, the precision of c-SN-MP outperformed SN-MP where the recall values are higher while the precision of SN-MP is higher for lower recall values. Based on the AUC of the PRD curves shown in Table 1, SN-MP and c-SN-MP patches performed similarly. In

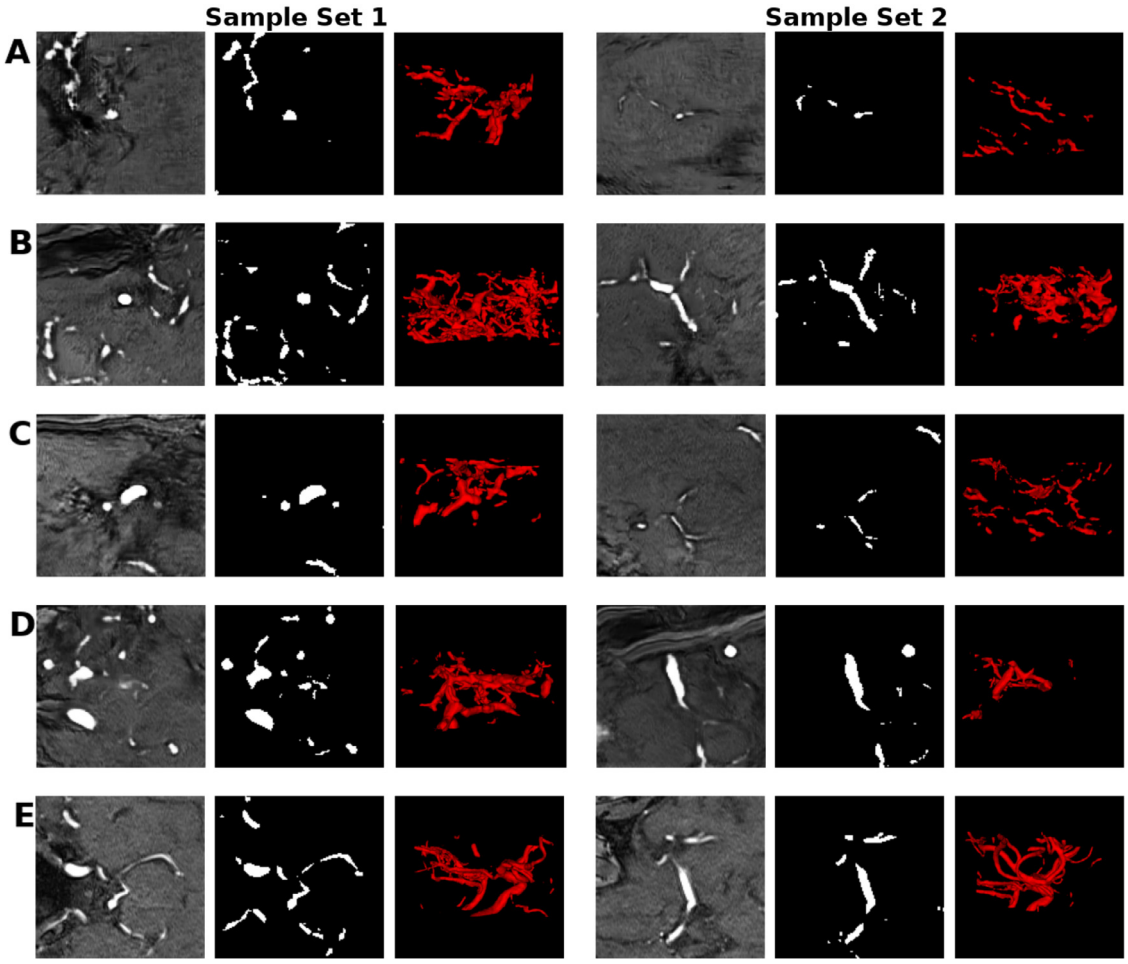

Fig. 4. Sets of samples of the mid-axial slice of the patch and label, and the corresponding 3D vessel structure from A) GP B) SN C) SN-MP D) c-SN-MP and E) real. The visualizations were obtained using ITK-SNAP for illustrative purposes only.

Table 2
Total number of trainable parameters, memory consumption and training times of various 3D GAN models. Note that c-SN-MP, which is our complex mixed precision model, uses twice the number of filters per layer leading to doubling of the trainable parameters compared to non-complex models. The memory consumption increased by 1.5 times compared to the SN model allowing it to be accommodated in the limited memory of our computational infrastructure. The training time also increased by 2.5 times but it was not a constraint in our study.

Model	Trainable parameters (million)	Memory (MB)	Time (hours)
GP model	145	15,085	78
SN model	145	14,333	77
SN-MP model	145	9,013	77
c-SN-MP model	308	21,351	192

Table 3
The mean DSC and mean bAVD (in voxels) across all the patients in the test set for 2 different datasets PEGASUS and 1000Plus. The value in brackets is the standard deviation across patients. A) All vessels is done on the entire prediction with the entire ground truth as reference, and B) Intracranial vessels is done on skull-stripped prediction with skull-stripped ground truth as reference.

Data source	PEGASUS		1000Plus	
	Mean DSC	Mean bAVD	Mean DSC	Mean bAVD
A) All vessels				
GP model	0.793 (0.024)	2.648 (1.189)	0.807 (0.03)	1.895(1.061)
SN model	0.804 (0.019)	2.425 (1.505)	0.796 (0.029)	1.855 (0.929)
SN-MP model	0.782 (0.020)	2.334 (1.122)	0.778 (0.032)	1.746 (0.894)
c-SN-MP model	***0.820 (0.017)***	***1.859 (1.038)***	***0.809 (0.031)***	***0.858 (0.91)***
Real	**0.906 (0.016)**	**0.339 (0.139)**	**0.883 (0.023)**	**0.554 (0.221)**
B) Intracranial vessels				
GP model	0.827 (0.015)	0.639 (0.132)	***0.829 (0.019)***	0.701 (0.195)
SN model	0.833 (0.013)	0.606 (0.141)	0.811 (0.023)	0.716 (0.213)
SN-MP model	0.804 (0.020)	0.784 (0.125)	0.785 (0.027)	0.822 (0.211)
c-SN-MP model	***0.841 (0.016)***	***0.508 (0.083)***	0.817 (0.028)	***0.611 (0.18)***
Real	**0.901 (0.019)**	**0.294 (0.077)**	**0.880 (0.024)**	**0.507 (0.126)**

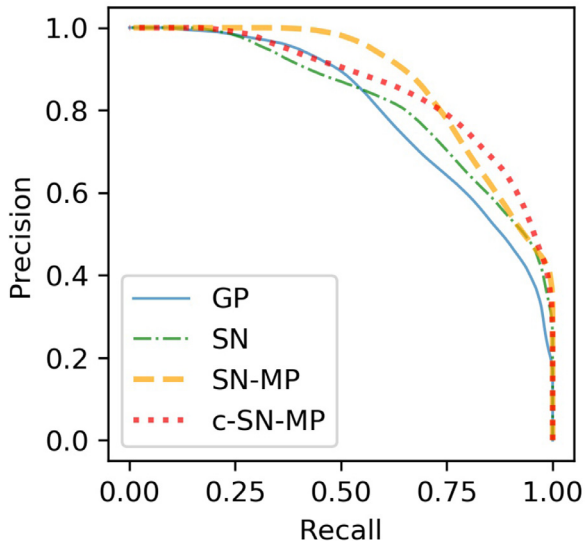

Fig. 5. PRD Curves of synthetic data from the four different models with real data as reference. Precision and Recall in GANs quantify the quality and modes captured by the models respectively.

Table 2, the memory consumption and the training duration of each of the GAN variants is shown. Using mixed precision improved the memory efficiency by approximately 40%.

The test set performance of the 3D U-Net trained on generated data from different models and on real data is shown in Table 3. Here, Table 3A shows the performance when all vessels are considered. The U-Net trained with c-SN-MP synthetic data outperformed all the U-Nets trained on other synthetic data for the PEGASUS test set (mean DSC 0.820; mean bAVD 1.859). In the case of the external dataset 1000Plus, the performance of U-Net trained on synthetic data from GP model and c-SN-MP model were the same in terms of mean DSC with 0.810 whereas the performance of U-Net trained on data from c-SN-MP was the lowest in terms of mean bAVD with 1.301. In comparison, the performance of the 3D U-Net trained with real data on PEGASUS test set was overall still the highest (mean DSC 0.906; mean bAVD 0.339), and on 1000Plus test set (mean DSC 0.887; mean bAVD 0.622).

Next, Table 3B shows the performance for intracranial vessels alone. Here, the U-Net trained with c-SN-MP synthetic data outperformed all the U-Nets trained on other synthetic data for the PEGASUS test set (mean DSC 0.841; mean bAVD 0.508). For the external test set from the 1000Plus dataset, the U-Net trained on generated data from GP was the highest in terms of mean DSC with 0.830 whereas the U-Net trained on generated data from c-SN-MP was the lowest in terms of mean bAVD with 0.639. The performance of labels with only intracranial vessels from the 3D U-Net trained with real data on the PEGASUS test set was still the highest (mean DSC 0.901; mean bAVD 0.294), and on the 1000Plus test set (mean DSC 0.880; mean bAVD 0.541).

Box-whisker plots of the prediction performance of various models on the two test sets are plotted in Fig. 6 which shows the inter-patient spread in performances for all vessels (Fig. 6A) and for intracranial vessels (Fig. 6B). The error maps of segmentation of two example patients, one from each of the two datasets, are shown in Fig. 7 for all vessels and for intracranial vessels.

4. Discussion

To the best of our knowledge, this is the first work to present generative adversarial network models that generate realistic 3D

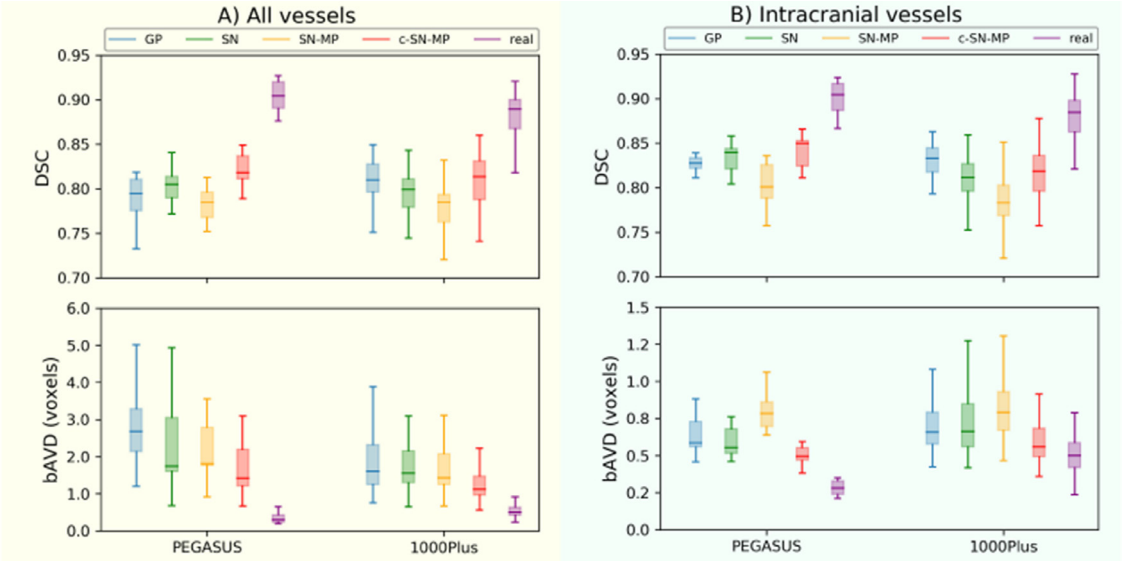

Fig. 6. Segmentation performance (DSC and bAVD) of 3D U-Net models trained with 4 different generated data and PEGASUS training data on the 2 datasets PEGASUS and 1000Plus of A) all vessels B) intracranial vessels. The horizontal line of the box-whisker plots indicates the median, the box indicates the interquartile range and the whiskers the minimum and maximum.

TOF-MRA volumes along with segmentation labels in medical imaging. We showed that utilizing mixed precision aids in achieving the highest image quality of synthetic data. Additionally, the synthetic data from our complex model maintained a substantial amount of predictive properties of the original volumes reflected by the good segmentation performance on real test data. These findings also held true on a second, independent dataset. The results showcase the potential of utilizing memory efficiency provided by mixed precision in designing a complex architecture. Increasing the complexity is required in order to generate high-resolution fine grained structures such as brain vessels in 3D TOF-MRA volumes from noise along with the corresponding segmentation labels. This work sets an important step towards sharing labeled 3D medical images that would facilitate better research in the medical imaging domain.

The segmentation performance of the 3D U-Net trained on synthetic data from our complex mixed precision model, c-SN-MP, showed the best performance compared to the other models based on synthetic data in terms of both metrics DSC and bAVD. Here, doubling the filters per layer in the GAN architecture is likely to have helped to capture the vessel structure in the training data. Also, visually it can be seen in Fig. 4 that c-SN-MP labels (Fig. 4D) look connected and most similar to real vessel structures (Fig. 4E). On the contrary, the labels of synthetic data from the simpler mixed precision model, SN-MP (Fig. 4C), are sparsely connected which explains the worst performance in terms of the DSC of the U-Net trained on SN-MP data. This seems plausible as the vessels are more relevant for segmentation than the background. The same can be observed in patch-label pairs from our most basic model, GP. Here, the segmentation performance was better than SN-MP in terms of DSC even though visually Fig. 4A shows that patch quality of GP is not as sharp as the other generated images.

In terms of quantitative measures of patch quality, the FID scores and PRD curves, the mixed precision models, both simple and complex, were rated to be of much better quality and variety when compared to models not using mixed precision (GP and SN models). However, the U-Net trained with the simpler mixed precision model, SN-MP, patch-label pairs had the lowest segmentation performance. A possible reason for this could be that FID and PRD curves, which are based on the features extracted only from the patches, might focus not only on the vessel structure but also the quality of the background. In contrast, the U-Net performance is more focused on recognizing the vessel structure. This is confirmed when looking at Fig. 4C where the patches seem realistic, but the vessel structures look disconnected. We see the reverse of this in the case of GP, where the patches look less realistic, but the vessel structures look more connected. This could explain why the GP model fared poorly in FID and PRD curves and yet did well in segmentation when used to train a U-Net. Looking more closely at the PRD curves (Fig. 5), the simpler mixed precision model, SN-MP, patches had good quality at lower recall values, while patches from our complex mixed precision model, c-SN-MP, had better quality when the recall values increased. This implies that c-SN-MP is capable of generating patches of slightly reduced quality but with higher variety, and thus, is better at handling mode collapse which is indicated by recall. While the FID and PRD curve provide insights regarding the image quality and variety, these metrics do not necessarily align with the performance in the vessel segmentation task. This emphasizes the importance of generating labels along with the image to determine the best generated data for the specific use case.

Overall, the FID and PRD curves indicated that more regularizations have a positive effect on the image quality and variety. The mixed precision models, SN-MP and c-SN-MP are the best performing models in terms of these metrics. They are both regularized with gradient penalty (Gulrajani et al., 2017) and spectral normalization (Miyato et al., 2018). These methods have been individually proposed to bound the critic by ensuring Lipschitz continuity which has been found to stabilize GAN training. Gradient penalty does this by applying a gradient based constraint to the objective function of the critic. With spectral normalization, the critic is bound by directly constraining its weight matrices by normalizing them with their spectral norm. Using the two methods together was proposed to be beneficial in the study that introduced spectral normalization in GANs (Miyato et al., 2018) and using them together has been shown to improve performance in another study (Kossen et al., 2021). In addition to these methods, we also used mixed precision for memory efficiency in the case of SN-MP and c-SN-MP models. Mixed precision has been found to act as yet another form of regularization (Micikevicius et al., 2018). Unlike FID and PRD curves, the segmentation performance does not always benefit from synthetic data generated by more regularized mod-

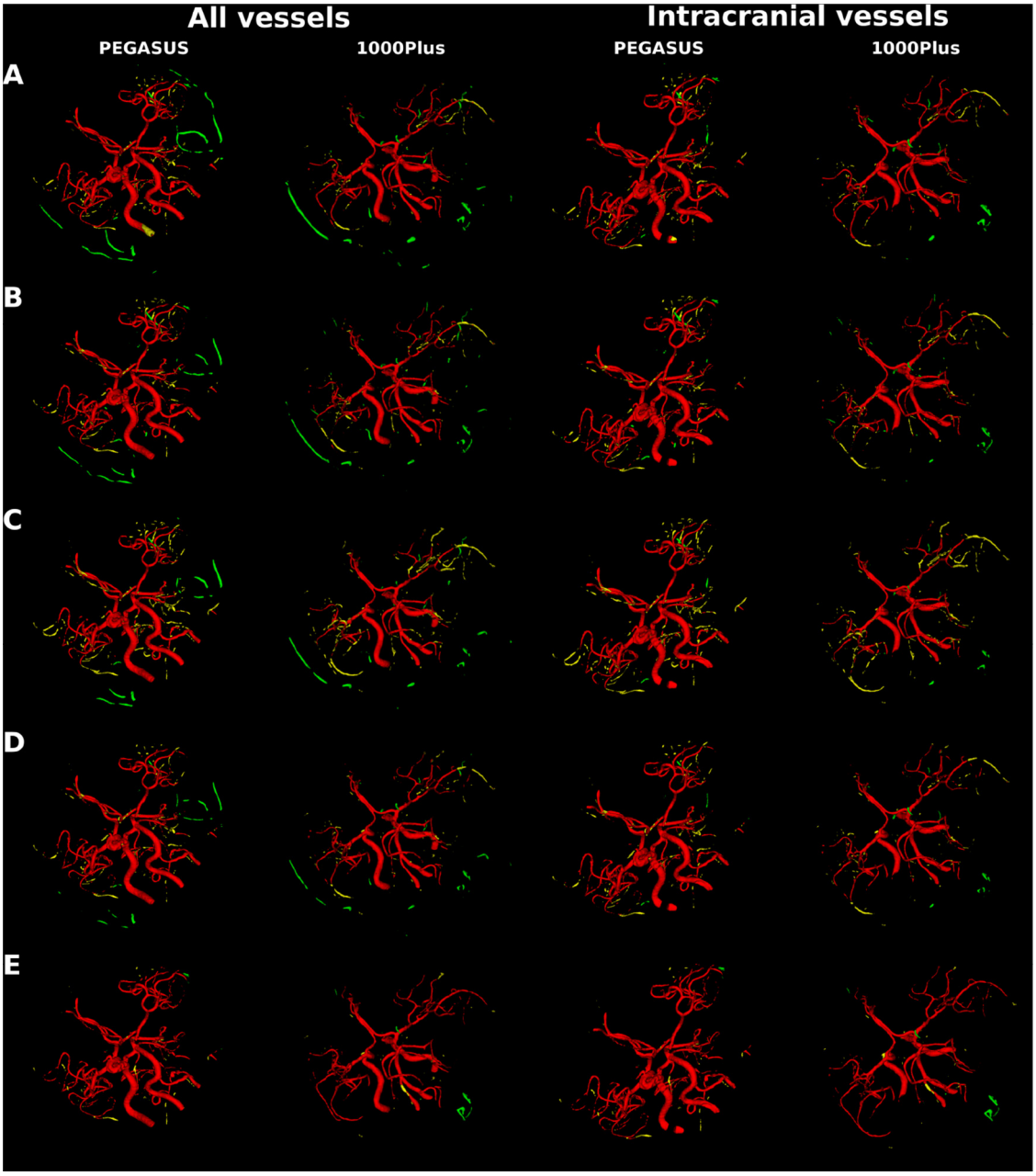

Fig. 7. Segmentation error map of an example patient each from PEGASUS test set and 1000Plus test set for all vessels and for intracranial vessels. Top to bottom maps from 3D U-Net model trained on: A. GP synthetic data B. SN synthetic data C. SN-MP synthetic data D. c-SN-MP synthetic data E. real data. True positives are shown in red, false positives are in green and false negatives in yellow.

els. When looking at the test DSC and bAVD of the U-Nets trained on synthetic data from the simpler models, GP, SN and SN-MP, it is difficult to rank the overall performance due to the varied differences in the two segmentation performance metrics across the two test sets. One possible explanation for the differences might be that regularization might positively impact the patches but not necessarily the binary segmentation labels that are much simpler to generate. To draw conclusions on how regularizations in GANs affect the two segmentation metrics and the generalizability to the additional set, a more systematic analysis would be required in further research. For the c-SN-MP model, we increased the model complexity which could better utilize the multiple regularizations and thus showed good segmentation performance as well as good image quality. An additional argument in favor of multiple regularizations is that it has been found to make models less vulnerable towards membership inference attacks (Truex et al., 2019; Chen et al., 2020). Such attacks are used by malicious parties to find out if a particular patient's data was used to train a model (Shokri et al., 2017). This is crucial to consider when sharing the synthetic data or the generator model. While regularization has been found useful to mitigate some attacks, applying differential privacy (DP) (Dwork and Roth, 2014) to the training process, by construction, puts an upper bound on the privacy leakage of the training data. DP is challenging to implement especially in a 3D GAN architecture, as it introduces a substantial number of parameters to an already overwhelming amount of parameters. This leads to high computational cost in terms of both memory and prolonged training time while reducing the test performance consider-

ably. To showcase this, we have included preliminary results with DP using Rényi divergence (Mironov, 2017) in 3D on a simplified GAN architecture in the appendix.

The U-Net trained on real data still outperforms all those trained on generated data. Here, Fig. 7 (All vessels) shows that all U-Net models trained with synthetic patches also segment blood vessels that are not brain blood vessels but rather vessels in the face and neck area, i.e. false positives. In contrast, the U-Net trained on real patches of the same size recognized if a vessel belonged to the brain and did not segment vessels outside the brain. This highlights that while the GANs learned to segment blood vessels, they did not learn to take the anatomical context of the vessels into account, i.e. the resulting models could not differentiate between face, neck and brain blood vessels. A possible explanation for this could be that the loss function for GANs focuses on the quality of the generated data and not the segmentation performance of the generated patch-label pairs compared to real patch-label pairs. A potential solution would be to generate labels with a first GAN similar in distribution to the real ground truth segmentation labels, and then a second GAN could be used to generate the corresponding image in inference mode with a 3D image-to-image translation GAN architecture that was trained with real labels and the corresponding images. However, training two 3D GANs separately would increase the overall training time substantially and was not feasible with our hardware infrastructure. We utilized an alternative solution where we applied the test image brain mask in a post-processing step leading to the removal of face and neck vessels. This is a valid post-processing approach since many clinical use cases only require segmentation of intracranial vessels. The performance of the 3D U-Net trained with generated data then improved, bringing it closer to the performance of the U-Net trained on real data as shown in Table 3B, Figs. 6B, 7 (Intracranial vessels).

Generating 3D data is more complex and computationally expensive compared to 2D. Yet, the best performing U-Net model trained on synthesized 3D data (DSC 0.841) is comparable to the best performing U-Net model trained on synthesized 2D data (DSC 0.848) for the same use case of intracranial vessel segmentation (Kossen et al., 2021). Here, the number of voxels that are generated is increased by a factor of 100 approximately. Meanwhile, only quarter of the number of filters per layer were used for all our non-complex 3D GAN models owing to memory limitations. In order to double the number of filters per layer for our complex model (c-SN-MP), we used mixed precision. Next, we also used upsampling instead of convtranspose to alleviate the checkerboard artifacts which increased the memory consumption substantially. Additionally, the training of WGAN requires the discriminator to be updated more often than the generator. Since this would lead to much longer training times, the current work utilized the Two Timescale Update Rule (TTUR). Here, the learning rate of the discriminator is set to be higher than that of the generator. These changes were crucial to cope with the special challenges of synthesis in 3D. Even with these restrictions, a similar segmentation performance of 3D in comparison with 2D underlines the importance of generating data in 3D to capture the contextual information within the third dimension for this 3D use case. It is likely that the segmentation performance of the U-Net trained with generated 3D data could surpass the performance of 2D data with more computational capacity, when more filters can be utilized in the 3D GAN architecture.

A different strategy with regards to data privacy is Federated Learning (FL). Here, sharing of data is avoided by locally computing updates for a global model that is then aggregated to be utilized by the participating clients. The results thus far are promising. However, standard FL does not create new data that can be made public for other research groups to access and improve model architectures. This is especially important in the case of rare pathologies where the data is scarce. Here, GANs can be used to generate data of such pathologies by research groups that have access to the data which can then be made publicly available. Additionally, there are technical and collaborative hurdles in FL such as picking a model-aggregation policy, standardization of hardware and software across multiple organizations among others (Ng et al., 2021). These challenges are more acute in the case of deep learning research. The organizational and collaborative efforts involved might not be feasible for research groups with limited resources. Since synthetic data from GANs can be shared, it provides easy and equitable access to all research groups investigating deep learning in medical imaging. FL, on the other hand, is more suitable for clinical application of well-established architectures with distributed training. It should be noted that both FL and GANs are susceptible to information leakage from the model weights even if the real data itself is not shared. This makes both methods open to privacy threats (Sheller et al., 2020; Chen et al., 2020). Here, DPGAN has been found useful (Xie et al., 2018). DP algorithms incorporate random noise into the model making them resilient towards information leakage (Shokri et al., 2017). FL and DPGANs could be taken together to combine their strengths as was done in FedDP-GAN (Zhang et al., 2021). In our work, we focus on the challenges of generating 3D medical imaging along with corresponding labels since labeling generated images is time and labour intensive. This is an important step before inclusion of DP into the GAN architecture. We have provided preliminary results using DP on a simple 3DGAN architecture in the appendix.

The main limitations of our study are computational in nature. First, we have not employed DP in the presented GAN architectures which would provide an upper bound on the information leakage when the generated data and/or generated model is shared. The computational load resulting from applying DP would have made the study unfeasible with the available computing infrastructure. Second, we did not use more novel GAN architectures validated on natural images such as Progressive GANs (Karras et al., 2018) or Multi-Scale-Gradients GANs (Karnewar and Wang, 2020). This is because of the multi-fold computational requirements of these architectures, especially in 3D. Patches of much smaller size could still be generated (Eklund, 2020), but they would not be very useful for the downstream task of vessel segmentation. Third, we generated patch-labels pairs and not whole volume-label pairs due to computational limitations. While a recently introduced hierarchical memory-efficient approach (Sun et al., 2021) might help to overcome the computational constraints, this would come at the cost of much longer training times considering 2 GANs of different resolutions are trained along with encoders in an end-to-end manner. Additionally, architectures that use data reconstruction are more susceptible to membership inference attack (Chen et al., 2020). Two of the recent studies (Kwon et al., 2019; Sun et al., 2021) generating 3D images alone use encoders in their architectures which make them less useful for the purpose of privacy-preserving data sharing. Lastly, we trained and tested our GAN architectures on one imaging modality, i.e. TOF-MRA. While we expect generalization of our results to other modalities that may not be high contrast-to-noise modalities like TOF-MRA, this should be verified in future studies. For that, we encourage other researchers to utilize our publicly available code. Our findings for TOF-MRA can be regarded as a first proof-of-concept that GAN architectures are able to synthesize realistic looking 3D volumes with corresponding segmentation labels.

5. Conclusion

In this study, we generated high resolution TOF-MRA patches along with their corresponding labels in 3D employing mixed precision for memory efficiency. Since most medical imaging is

recorded in 3D, generating 3D images that retain the volumetric information together with labels that are time-intensive to generate manually is a first step towards sharing labeled data. While our approach is not privacy-preserving yet, the architecture was designed with privacy as a key aspiration. It would be possible to extend it with differential privacy in future works once the computational advancements allow it. This would pave the way for sharing privacy-preserving, labeled 3D imaging data. Research groups could utilize our open source code to implement a mixed precision approach to generate 3D synthetic volumes and labels efficiently and verify if they hold the necessary predictive properties for the specific downstream task. Making such synthetic data available on request would then allow for larger heterogeneous datasets to be used in the future alleviating the typical data shortages in this domain. This will pave the way for robust and replicable model development and will facilitate clinical applications.

Declaration of Competing Interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

None of the authors have direct or indirect financial interest in the subject matter discussed in the manuscript.

However, the following disclosures unrelated to the current work is as follows:

Pooja Subramaniam reported receiving personal fees from ai4medicine outside the submitted work. Tabea Kossen reported receiving personal fees from ai4medicine outside the submitted work. Dr Madai reported receiving personal fees from ai4medicine outside the submitted work. Adam Hilbert reported receiving personal fees from ai4medicine outside the submitted work. Dr Frey reported receiving grants from the European Commission, reported receiving personal fees from and holding an equity interest in ai4medicine outside the submitted work. There is no connection, commercial exploitation, transfer or association between the projects of ai4medicine and the results presented in this work. While not related to this work, Dr Sobesky reports receipt of speakers honoraria from Pfizer, Boehringer Ingelheim, and Daiichi Sankyo. Furthermore, Dr Fiebach has received consulting and advisory board fees from BioClinica, Cerevast, Artemida, Brainomix, Biogen, BMS, EISAI, and Guerbet.

CRediT authorship contribution statement

Pooja Subramaniam: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tabea Kossen:** Conceptualization, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Kerstin Ritter:** Supervision, Writing – review & editing. **Anja Hennemuth:** Supervision, Writing – review & editing. **Kristian Hildebrand:** Supervision, Writing – review & editing. **Adam Hilbert:** Conceptualization, Writing – review & editing. **Jan Sobesky:** Data curation, Writing – review & editing. **Michelle Livne:** Conceptualization, Writing – review & editing. **Ivana Galinovic:** Data curation, Writing – review & editing. **Ahmed A. Khalil:** Data curation, Writing – review & editing. **Jochen B. Fiebach:** Data curation, Writing – review & editing. **Dietmar Frey:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Vince I. Madai:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

This work has received funding by the German Federal Ministry of Education and Research through (1) the grant Centre for Stroke Research Berlin and (2) a Go-Bio grant for the research group PREDICTioN2020 (lead: DF). Grant number 031B0154.

Appendix A. Data augmentation

An additional analysis to complement the evaluation of our synthetic data is to use it to augment the training data for the downstream brain blood vessel segmentation task. Here we trained segmentation models with PEGASUS training data and augmented it with synthetic data from the 4 GAN models separately for additional analysis. Table A.1 summarizes the segmentation results for all vessels (Table A.1 A) and intracranial vessels (Table A.1B). Fig. A.1 is a box-whisker plot to visualize the spread in the seg-

Table A.1
The mean DSC and mean bAVD (in voxels) across all the patients in the test set for 2 different datasets PEGASUS and 1000Plus using model trained with real data along with generated data used as data augmentation. The value in brackets is the standard deviation across patients. A) All vessels is done on the entire prediction with the entire ground truth as reference, and B) Intracranial vessels is done on skull-stripped prediction with skull-stripped ground truth as reference.

Data source	PEGASUS		1000Plus	
	Mean DSC	Mean bAVD	Mean DSC	Mean bAVD
A) All vessels				
Real + GP model	0.902 (0.046)	**0.333 (0.151)**	0.862 (0.029)	0.65 (0.271)
Real + SN model	0.906 (0.016)	0.385 (0.145)	0.878 (0.021)	0.558 (0.199)
Real + SN-MP model	0.903 (0.013)	0.359 (0.133)	0.883 (0.02)	**0.511 (0.145)**
Real + c-SN-MP model	**0.907 (0.012)**	0.399 (0.204)	**0.891 (0.02)**	0.564 (0.222)
Real	0.906 (0.016)	0.339 (0.139)	0.883 (0.023)	0.554 (0.221)
B) Intracranial vessels				
Real + GP model	0.897 (0.018)	0.323 (0.073)	0.855 (0.029)	0.626 (0.199)
Real + SN model	0.905 (0.017)	0.318 (0.075)	0.874 (0.022)	0.546 (0.148)
Real + SN-MP model	0.900 (0.017)	0.328 (0.078)	0.877 (0.02)	0.518 (0.121)
Real + c-SN-MP model	**0.905 (0.016)**	0.306 (0.091)	**0.884 (0.02)**	0.557 (0.129)
Real	0.901 (0.019)	**0.294 (0.077)**	0.880 (0.024)	**0.507 (0.126)**

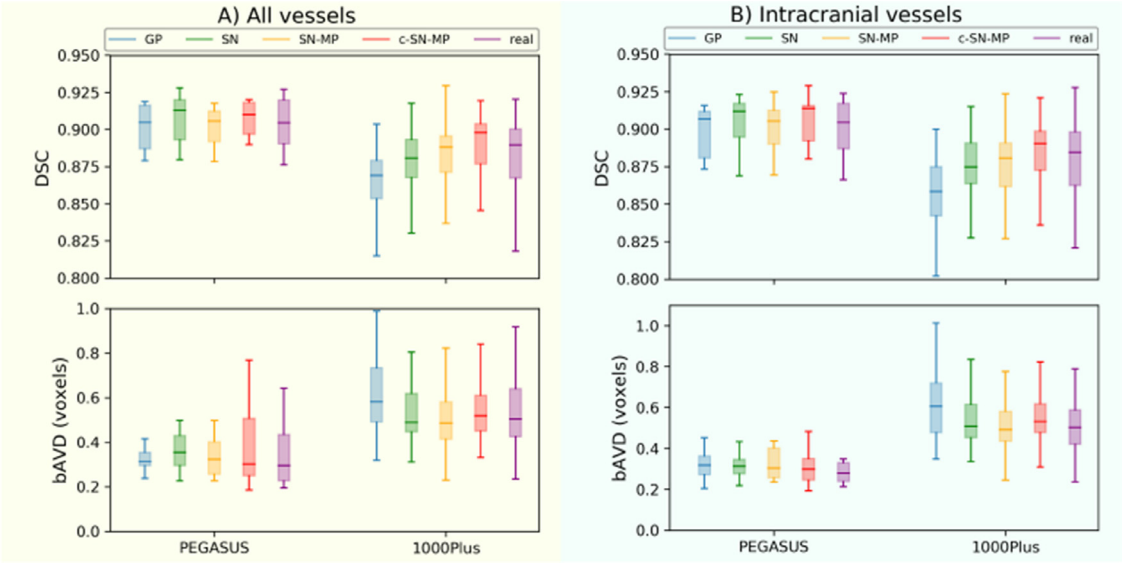

Fig. A.1. Segmentation performance (DSC and bAVD) of 3D U-Net models trained with PEGASUS training data together with 4 different generated data as data augmentation on the 2 datasets PEGASUS and 1000Plus of A) all vessels B) intracranial vessels. The horizontal line of the box-whisker plots indicates the median, the box indicates the interquartile range and the whiskers the minimum and maximum.

mentation performance between patients for all vessels (Fig. A.1A) and intracranial vessels (Fig. A.1B).

Using synthetic data from c-SN-MP model to augment the real data for training segmentation model provided slightly better mean DSC on both test sets (PEGASUS and 1000Plus) for the two cases of A) all vessels and B) intracranial vessels when compared to using only real data or using synthetic data from other GAN models along with real data. While data augmentation is a valid application of our synthetic data, the additional value from them is limited as can be seen from the results. This could be because the predictive properties captured by the synthesized data is similar to the real data. This was also the case in the study with 2D GAN (Kossen et al., 2021) where data augmentation with 2D generated data did not lead to substantial difference in the segmentation performance.

Appendix B. 3D differentially private GAN

Differential privacy (DP) is a natural mitigation strategy against membership inference threats. Using DP to synthesize data would allow accounting of the level of possible re-identification thus providing privacy guarantees of the generated data. In order to illus-

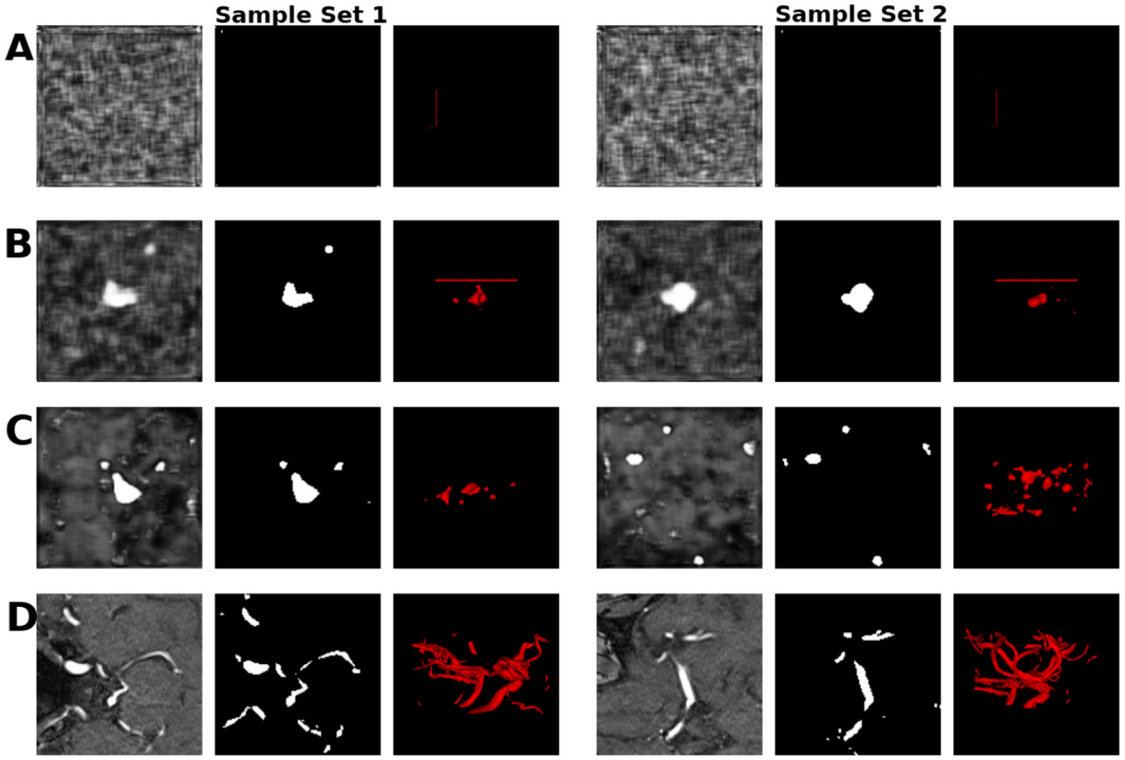

Fig. B.1. Sets of samples of the mid-axial slice of the patch and label, and the corresponding 3D vessel structure from A) DPGAN $\epsilon \approx 10^2$ B) DPGAN $\epsilon \approx 10^3$ C) DPGAN $\epsilon \approx 10^6$ D) real. Note that lower the ϵ higher the privacy. The visualizations were obtained using ITK-SNAP for illustrative purposes only.

Table B.1
The mean DSC and mean bAVD (in voxels) across all the patients in the test set for 2 different datasets PEGASUS and 1000Plus using model trained with generated data from 3D DPGAN with different ϵ values - starting from low ϵ value indicating high privacy to the high ϵ value indicating low privacy. The value in brackets is the standard deviation across patients. A) All vessels is done on the entire prediction with the entire ground truth as reference, and B) Intracranial vessels is done on skull-stripped prediction with skull-stripped ground truth as reference.

Data source	PEGASUS		1000Plus	
	Mean DSC	Mean bAVD	Mean DSC	Mean bAVD
A) All vessels				
DPGAN $\epsilon \approx 10^2$	0.085 (0.012)	***4.509 (0.903)***	0.083 (0.016)	4.116 (0.743)
DPGAN $\epsilon \approx 10^3$	0.562 (0.050)	6.307 (2.406)	0.567 (0.041)	4.624 (1.608)
DPGAN $\epsilon \approx 10^6$	***0.581 (0.048)***	5.05 (2.267)	***0.568 (0.041)***	***3.962 (1.591)***
Real	**0.906 (0.016)**	**0.339 (0.139)**	**0.883 (0.023)**	**0.554 (0.221)**
B) Intracranial vessels				
DPGAN $\epsilon \approx 10^2$	0.081 (0.013)	4.77 (1.081)	0.077 (0.015)	4.595 (0.878)
DPGAN $\epsilon \approx 10^3$	0.586 (0.045)	3.141 (0.514)	0.569 (0.045)	2.698 (0.604)
DPGAN $\epsilon \approx 10^6$	***0.604 (0.048)***	***2.201 (0.413)***	***0.572 (0.048)***	***2.001 (0.445)***
Real	**0.901 (0.019)**	**0.294 (0.077)**	**0.88 (0.024)**	**0.507 (0.126)**

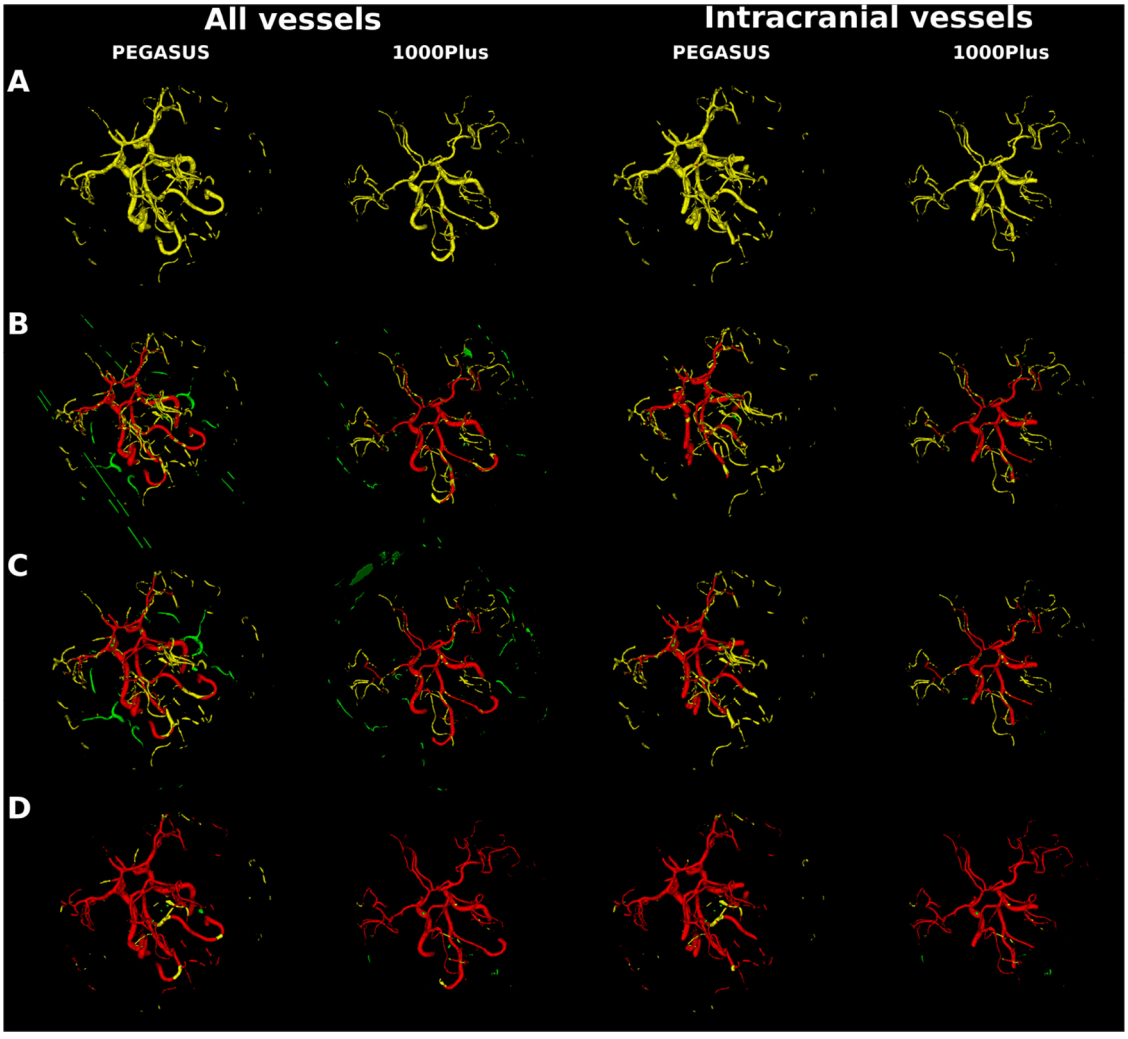

Fig. B.2. Segmentation error map of an example patient each from PEGASUS test set and 1000Plus test set for all vessels and for intracranial vessels. Top to bottom maps from 3D U-Net model trained on: A. DPGAN $\epsilon \approx 10^2$ B. DPGAN $\epsilon \approx 10^3$ C. DPGAN $\epsilon \approx 10^6$ D. real data. True positives are shown in red, false positives are in green and false negatives in yellow.

trate this, we utilized the Opacus package from PyTorch to apply DP-SGD algorithm with Rényi divergence on a 3D adapted version of the WGAN (Arjovsky et al., 2017). Henceforth, we refer to this model architecture as 3D DPGAN. Here, we clipped the weights of the critic with a clipping parameter of 0.01. We had to halve the number of filters per layer in both the critic and the generator in order to be able to train the 3D DPGAN within our computational infrastructure. We trained the 3D WGAN with Rényi differential privacy accountant which translates to (ϵ, δ)-DP guarantees. The (ϵ, δ) pairs quantify the privacy properties of DP-SGD. ϵ is the measure of privacy loss at a differential change in data with δ probability that the privacy constraint of ϵ does not hold true. A smaller ϵ value leads to better privacy. For comparing synthetic data with different privacy guarantees, the noise multiplier values were set to different values [0.1, 0.3, 0.5] which each provide ϵ values $[\approx 10^6, 10^3, 10^2]$ respectively. It should be noted that since the training samples consist of 3D patch-label pairs from the TOF-MRA image-segmentation label pairs, the guarantees showcased here also pertain to the patch-label pair data rather than the whole TOF-MRA image of a patient. We set δ to the inverse of the number of training samples following convention (Torkzadehmahani et al., 2019). The maximum gradient norm value of 1 was applied for clipping gradients. Adam optimizer with a learning rate of 0.0001 for both the critic and the generator was used instead of the TTUR method as the training time was reasonable with 5 updates of critic for every update of the generator. All the GANs were trained for 100 epochs. The threshold of 0.7 was applied on the generated labels from DPGAN $\epsilon \approx 10^6$ and 0.6 for DPGAN $\epsilon \approx 10^2$ and $\epsilon \approx 10^3$ chosen based on the segmentation performance on the validation set. The code for the same is also made available in the GitHub repository that has already been provided.

The generated patch-label pairs and the 3D vessel structure synthesized with different ϵ values are shown in Fig. B.1 along with the real patch-label pairs for a qualitative comparison. With decreasing ϵ values the generated data quality reduces. In other words, higher privacy guarantees come with lower quality. This is also supported quantitatively with lower segmentation performance of those U-Nets trained with generated data from lower ϵ DPGAN and vice-versa. Table B.1 shows the results of the test segmentation performance on 2 datasets trained with generated patch-label pairs from DPGANs with different ϵ values for A) all vessels and B) intracranial vessels only. Synthetic data used from DPGAN with $\epsilon \approx 10^6$ has the best performance in the case of all vessels (mean DSC 0.581) and in the case of intracranial vessels (mean DSC 0.604; mean bAVD 2.201). bAVD of U-Net trained with synthetic data from DPGAN with $\epsilon \approx 10^2$ is unexpectedly lower (mean bAVD 4.509) than that trained with $\epsilon \approx 10^6$ (mean bAVD 5.05). This is because the metric bAVD penalizes false positives more than false negatives. This explanation is corroborated in Fig. B.2 which visualizes the error masks of segmentation of two example patients, one from each of the two datasets for all vessels and intracranial vessels. Fig. B.2A (PEGASUS) - All vessels shows the segmentation error maps from U-Net trained on synthetic data from the highest privacy guarantee of $\epsilon \approx 10^2$. The network misses almost all the vessels and yet the bAVD is lower than bAVD of U-Net trained on data from DPGAN $\epsilon \approx 10^6$ (Fig. B.2C (PEGASUS) - All vessels) which has far less false negatives but relatively more false positives owing to vessels from neck and face area. This is further confirmed when these vessels are removed for analysis by the post-process skull-stripping of the labels. Then, the bAVD of U-Net trained with DPGAN $\epsilon \approx 10^6$ (mean bAVD 2.201) improves much more than that of U-Net trained with DPGAN $\epsilon \approx 10^2$ (mean bAVD 4.77).

Our results for the 3D DPGAN show that the generated data with the largest epsilon $\epsilon \approx 10^6$ yielded the best performance (mean DSC 0.604). While this model provided an upper bound of privacy, it should be noted that $\epsilon \approx 10^6$ is a very large value and the resulting privacy bounds are thus too loose. Moreover, the performance of our DPGAN with $\epsilon \approx 10^6$ is quite low compared to the performance of our generated data without any privacy guarantees (mean DSC 0.841). Therefore, we conclude that finding the right balance between privacy and utility remains a challenge for differential privacy to be used even in a very simple 3D GAN architecture.

Supplementary material

E-supplementary data of this work can be found in online version of the paper.

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102396.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv:1701.07875 [cs, stat].

Aydin, O. U., Taha, A. A., Hilbert, A., Khalil, A. A., Galinovic, I., Fiebach, J. B., Frey, D., Madai, V. I., 2021. An evaluation of performance measures for arterial brain vessel segmentation. Accepted for publication

Aydin, O.U., Taha, A.A., Hilbert, A., Khalil, A.A., Galinovic, I., Fiebach, J.B., Frey, D., Madai, V.I., 2021. On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. Eur. Radiol. Exp. 5 (1), 4. doi:10.1186/s41747-020-00200-2.

Baur, C., Albarqouni, S., Navab, N., 2018. Generating highly realistic images of skin lesions with GANs. arXiv:1809.01410 [cs, eess].

Bermudez, C., Plassard, A.J., Davis, T.L., Newton, A.T., Resnick, S.M., Landman, B.A., 2018. Learning implicit brain MRI manifolds with deep learning. Proc SPIE Int. Soc. Opt. Eng. 10574. doi:10.1117/12.2293515.

Chen, D., Yu, N., Zhang, Y., Fritz, M., 2020. GAN-leaks: a taxonomy of membership inference attacks against generative models. arXiv:1909.03935 [cs]. 10.1145/3372297.3417238

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: transfer learning for 3D medical image analysis. arXiv:1904.00625 [cs].

Clinical Practice Committee, 2000. Informed consent for medical photographs. Dysmorphology subcommittee of the clinical practice committee, american college of medical genetics. Genet. Med. 2 (6), 353–355. doi:10.1097/00125817-200011000-00010.

Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Foundations Trends Theor. Comput. Sci. 9 (3–4), 211–407. doi:10.1561/0400000042.

Eklund, A., 2020. Feeding the zombies: synthesizing brain volumes using a 3D progressive growing GAN. arXiv:1912.05357 [cs, eess].

Foroozandeh, M., Eklund, A., 2020. Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE. arXiv:2009.05946 [cs]version: 1.

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

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv:1406.2661 [cs, stat].

Greenspan, H., van Ginneken, B., Summers, R.M., 2016. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging 35 (5), 1153–1159. doi:10.1109/TMI.2016.2553401.

Guibas, J. T., Virdi, T. S., Li, P. S., 2018. Synthetic medical images from dual generative adversarial networks. arXiv:1709.01872 [cs]version: 3.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein GANs. arXiv:1704.00028 [cs, stat].

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2018. GANs trained by a two time-scale update rule converge to a local nash equilibrium. arXiv:1706.08500 [cs, stat].

Hilbert, A., Madai, V.I., Akay, E.M., Aydin, O.U., Behland, J., Sobesky, J., Galinovic, I., Khalil, A.A., Taha, A.A., W&uumlrfel, J., Dusek, P., Niendorf, T., Fiebach, J.B., Frey, D., Livne, M., 2020. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. Neurology doi:10.1101/2020.04.08.20057570. preprint

Hotter, B., Pittl, S., Ebinger, M., Oepen, G., Jegzentis, K., Kudo, K., Rozanski, M., Schmidt, W., Brunecker, P., Xu, C., Martus, P., Endres, M., Jungehülsing, G., Villringer, A., Fiebach, J., 2009. Prospective study on the mismatch concept in acute stroke patients within the first 24 h after symptom onset - 1000Plus study. BMC Neurol. 9, 60. doi:10.1186/1471-2377-9-60.

Karnewar, A., Wang, O., 2020. MSG-GAN: multi-scale gradients for generative adversarial networks. arXiv:1903.06048 [cs, stat].

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196 [cs, stat].

Kingma, D., Ba, J., 2014. Adam: a method for stochastic optimization. In: International Conference on Learning Representations.

Kossen, T., Subramaniam, P., Madai, V.I., Hennemuth, A., Hildebrand, K., Hilbert, A., Sobesky, J., Livne, M., Galinovic, I., Khalil, A.A., Fiebach, J.B., Frey, D., 2021. Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. Comput. Biol. Med. 131, 104254. doi:10.1016/j.compbiomed.2021.104254.

Kwon, G., Han, C., Kim, D., 2019. Generation of 3D brain MRI using auto-encoding generative adversarial networks. MICCAI doi:10.1007/978-3-030-32248-9_14.

Livne, M., Rieger, J., Aydin, O.U., Taha, A.A., Akay, E.M., Kossen, T., Sobesky, J., Kelleher, J.D., Hildebrand, K., Frey, D., Madai, V.I., 2019. A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. Front. Neurosci. 13. doi:10.3389/fnins.2019.00097.

Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 29 (2), 102–127. doi:10.1016/j.zemedi.2018.11.002.

Masoudi, S., Harmon, S.A.A., Mehralivand, S., Walker, S.M., Raviprakash, H., Bagci, U., Choyke, P.L., Turkbey, B., 2021. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. J. Med. Imaging 8 (1), 010901. doi:10.1117/1.JMI.8.1.010901.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H., 2018. Mixed precision training. arXiv:1710.03740 [cs, stat].

Mironov, I., 2017. Renyi differential privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pp. 263–275. doi:10.1109/CSF.2017.11.

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. arXiv:1802.05957 [cs, stat].

Mutke, M.A., Madai, V.I., von Samson-Himmelstjerna, F.C., Zaro Weber, O., Revankar, G.S., Martin, S.Z., Stengl, K.L., Bauer, M., Hetzer, S., Günther, M., Sobesky, J., 2014. Clinical evaluation of an arterial-spin-labeling product sequence in steno-occlusive disease of the brain. PLoS ONE 9 (2), e87143. doi:10.1371/journal.pone.0087143.

Neff, T., Payer, C., Återn, D., Urschler, M., 2018. Generative adversarial networks to synthetically augment data for deep learning based image segmentation. In: Proceedings of the OAGM Workshop 2018 doi:10.3217/978-3-85125-603-1-07.

Ng, D., Lan, X., Yao, M.M.-S., Chan, W.P., Feng, M., 2021. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant. Imaging Med. Surg. 11 (2), 852–857. doi:10.21037/qims-20-595.

Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., Gelly, S., 2018. Assessing generative models via precision and recall. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp. 5234–5243.

Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., Bakas, S., 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. 10 (1), 12598. doi:10.1038/s41598-020-69250-1.

Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership inference attacks against machine learning models. arXiv:1610.05820 [cs, stat].

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K., 2021. Hierarchical amortized training for memory-efficient high resolution 3D GAN. arXiv:2008.01910 [cs, eess].

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2016.308.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15. doi:10.1186/s12880-015-0068-x.

Torkzadehmahani, R., Kairouz, P., Paten, B., 2019. DP-CGAN: differentially private synthetic data and label generation. pp. 0–0 https://openaccess.thecvf.com/content_CVPRW_2019/html/CV-COPS/Torkzadehmahani_DP-CGAN_Differentially_Private_Synthetic_Data_and_Label_Generation_CVPRW_2019_paper.html.

Truex, S., Liu, L., Gursoy, M. E., Yu, L., Wei, W., 2019. Towards demystifying membership inference attacks. arXiv:1807.09173 [cs].

Valizadeh, S.A., Liem, F., Mérillat, S., Hänggi, J., Jäncke, L., 2018. Identification of individual subjects on the basis of their brain anatomical features. Sci. Rep. 8 (1), 5611. doi:10.1038/s41598-018-23696-6.

Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M., Alzheimer's Disease Neuroimaging Initiative, 2015. BrainPrint: a discriminative characterization of brain morphology. Neuroimage 109, 232–248. doi:10.1016/j.neuroimage.2015.01.032.

Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P., 2020. Preparing medical imaging data for machine learning. Radiology 295 (1), 4–15. doi:10.1148/radiol.2020192224.

Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J., 2018. Differentially private generative adversarial network. arXiv:1802.06739 [cs, stat].

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

Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., Wu, Y., 2021. FedDPGAN: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. Inf. Syst. Front. doi:10.1007/s10796-021-10144-6.