www.nature.com/scientificdata

scientific data

OPEN

ANALYSIS

Brain tumor segmentation using synthetic MR images - A comparison of GANs and diffusion models

Muhammad Usman Akbar[1,2], Måns Larsson[3], Ida Blystad[2,4] & Anders Eklund[1,2,5]✉

Large annotated datasets are required for training deep learning models, but in medical imaging data sharing is often complicated due to ethics, anonymization and data protection legislation. Generative AI models, such as generative adversarial networks (GANs) and diffusion models, can today produce very realistic synthetic images, and can potentially facilitate data sharing. However, in order to share synthetic medical images it must first be demonstrated that they can be used for training different networks with acceptable performance. Here, we therefore comprehensively evaluate four GANs (progressive GAN, StyleGAN 1–3) and a diffusion model for the task of brain tumor segmentation (using two segmentation networks, U-Net and a Swin transformer). Our results show that segmentation networks trained on synthetic images reach Dice scores that are 80%–90% of Dice scores when training with real images, but that memorization of the training images can be a problem for diffusion models if the original dataset is too small. Our conclusion is that sharing synthetic medical images is a viable option to sharing real images, but that further work is required. The trained generative models and the generated synthetic images are shared on AIDA data hub.

Introduction

Medical imaging plays a vital role in the diagnosis and treatment of many diseases, enabling healthcare professionals to understand and visualize the internal structures and functions of the human body. With the advancement of artificial intelligence (AI) the field of medical imaging has seen significant improvements in terms of accuracy, efficiency, and cost-effectiveness. AI techniques such as machine learning and deep learning are commonly applied to medical imaging to, for instance, facilitate early detection and diagnosis of diseases and speedup time consuming segmentations[1,2]. For example, radiotherapy treatment planning requires segmentation of the tumor and several organs at risk. It is still common that these segmentations are done manually, and segmentation networks can here be used to reduce the required time for one patient from hours to a few minutes[3].

However, training deep learning models, such as convolutional neural networks (CNNs) and vision transformers, for classification or segmentation normally requires large annotated datasets as the models may have millions of parameters. In computer vision, tremendous progress has been made during the last 10 years, and a crucial resource is the open ImageNet database[4] which contains more than 14 million labeled images. Techniques developed in computer vision are rapidly transferred to the medical imaging field, but a major constraint is that access to medical images is much more complicated due to ethics, anonymization and data protection legislation (e.g. the general data protection regulation (GDPR)). There are several openly available medical imaging datasets, but they are much smaller compared to ImageNet (for example, the human connectome project (HCP) shares 1,100 subjects[5], OpenNeuro shares about 30,000[6], UK biobank will scan and share 100,000[7]). Furthermore, openly available data are often anonymized through defacing, can represent selective populations around universities, focus on healthy controls rather than diseased populations, and are

[1]Department of Biomedical Engineering, Linköping University, Linköping, Sweden. [2]Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden. [3]Eigenvision, Malmö, Sweden. [4]Department of Radiology in Linköping and Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden. [5]Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, Linköping, Sweden. ✉e-mail: anders.eklund@liu.se

often curated before distribution to eliminate bad quality data. This limits the potential applicability of any model trained on such data in clinical settings. Hospitals have records containing immense quantities of medical images, but these records are often not accessible for research due to regulatory hurdles.

Generative models, such as generative adversarial networks (GANs) and diffusion models, can today produce very realistic synthetic images, by learning the high dimensional distribution of the training images. A potential solution to facilitate sharing of medical images is therefore to generate and share synthetic images, or more precisely synthetic patients, as GDPR should not apply to medical images which do not belong to a specific person (but further legal research is needed). Recent work has demonstrated that generative models (especially diffusion models) can memorize the training images[8–10], meaning that the synthetic images are just copies of the training images. As this questions the validity of sharing synthetic medical images, it is thoroughly discussed at the end of this paper.

To share synthetic medical images, and to motivate further research regarding legal aspects and memorization, it must first be demonstrated that they can be used for training deep learning models with acceptable performance. Due to the growing number of generative image models, one must also select the best model.

Related work. Rankin *et al.*[11] used 19 open health datasets to understand the difference in performance of supervised machine learning models trained on synthetic data compared with those trained on real data, but only used tabular datasets and no image data. Similarly, El Emam *et al.*[12] used synthetic tabular data from COVID-19 patients to predict death, and obtained similar performance using synthetic data. Using synthetic images for training CNNs for classification has become popular during recent years[13–18], especially in medical imaging[19] where obtaining large annotated datasets is much more time consuming compared to computer vision. On the other hand, related work on training segmentation networks with synthetic images and corresponding annotations is more limited. To generate synthetic images and the corresponding annotations can be done in at least two ways; jointly as a multi-channel image[20–22] or as a two-step process where one model generates a synthetic label (annotation) image and another model generates the medical image from the label image[23–26]. Bowles *et al.*[20] demonstrated that adding synthetic images from a 2D GAN lead to improvements of Dice similarity coefficient between 1 and 5 percent, but did not perform training with only synthetic images. Shin *et al.*[24] also demonstrated small improvements when adding synthetic images as augmentation. Thambawita *et al.*[27] compared different GANs for generating synthetic colonoscopy images and annotations, but did not use more recent models like StyleGAN or diffusion models. Fernandez *et al.*[28] used the two-step approach to generate label images with a diffusion model, and then used SPADE[29] to generate the medical image from the label image. They also applied their models to brain tumor images, but only performed a binary tumor segmentation and did not compare with any other generative models. Furthermore, the generative model was only trained with 1064 slices from 5 subjects.

This work. Here we comprehensively evaluate four 2D GANs (progressive GAN[30], StyleGAN 1–3[31–33]) and a 2D diffusion model[34,35] for generating brain tumor images and tumor annotations, using two openly available datasets (BraTS 2020 and 2021[36–41]). We demonstrate that using synthetic images for training segmentation networks (a U-Net[42] and a Swin transformer[43]) leads to performance metrics which are slightly lower compared to training with real images, and that sharing synthetic images therefore is a viable option to sharing real images (as long as one verifies that the synthetic images are not too similar to the training images). To the best of our knowledge, no such comprehensive evaluation, requiring more than 2000 GPU days for training all models, has previously been performed for medical image segmentation. The trained generative models and the generated synthetic images are shared on AIDA data hub[44,45].

Results

Figure 1 shows a real 5-channel image, and a randomly selected synthetic slice from each generative model. To investigate how distinct the synthetic images are from the training images, we calculated the highest correlation between 100 synthetic images and all training images (see our related work on memorization[10]). Briefly, synthetic images from a GAN show a distribution of highest correlations which is similar to when comparing training images and test images. For diffusion models, many of the synthetic images are very similar to a training image. Figures 2, 3 show the resulting U-Net segmentations for a random slice in the two test sets, when training the network with different settings.

Evaluation metrics. To compare the five generative models we use a variety of metrics. The quality and diversity of synthetic images are often evaluated using metrics such as Frechét inception distance (FID) and inception score (IS)[46], which use pre-trained CNNs to calculate how different the activations in the CNNs are when feeding real and synthetic images through them. The most important evaluation is in our opinion to train segmentation networks with the synthetic images, and then test how these networks perform on real images. Here we used a U-Net[42], as it is one of the most common networks for medical image segmentation, and a Swin transformer[43], to see how the results generalize to a more recent network, see the Methods section for details. Segmentation networks are normally evaluated using Dice (measures overlap between true and predicted annotations) and Hausdorff distance (measures the greatest of all the distances from a point in one set to the closest point in the other set). Augmentation is often applied when training segmentation networks, and the segmentation networks were therefore trained with and without augmentation (see Methods section for details).

Ranking of generative models. To summarize all the results, Table 1 shows the ranking of the five generative models, based on the different metrics FID, IS, Dice and Hausdorff distance. Here we focus on training segmentation networks with synthetic images only, to not make the table too complicated. The diffusion model performs best when comparing the models in terms of Dice and Hausdorff distance, but unfortunately this is in several cases explained by memorization. As expected, the older progressive GAN model often performs worse

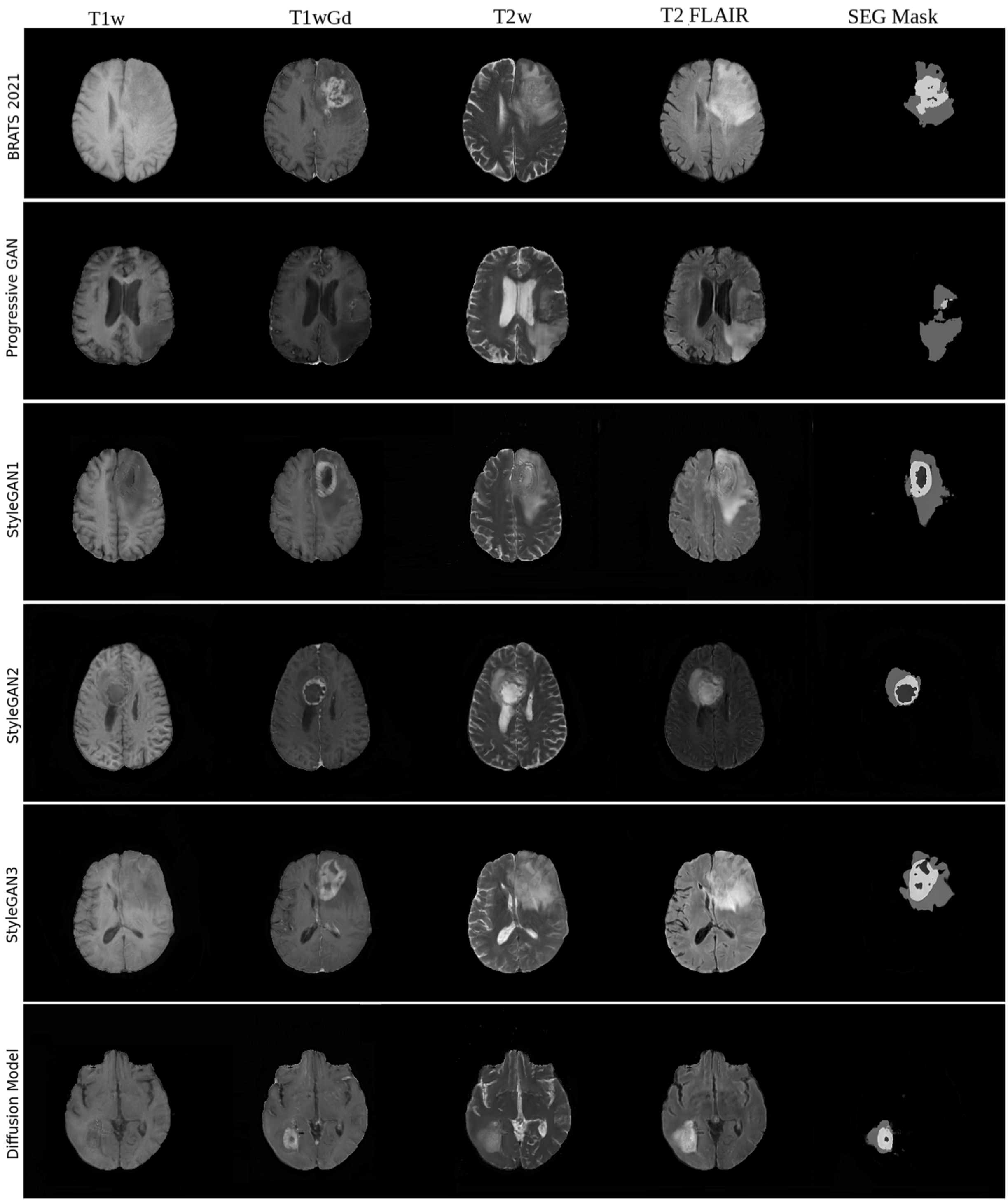

Fig. 1 Synthetic 5-channel images from the BraTS 2021 data. Each row shows a generative model, except for the top row which shows a real example, and each column shows a different MR sequence.

compared to more recent StyleGAN models. Overall the rankings are similar for U-Net and the Swin transformer. Clearly, the rankings according to the common FID and IS metrics (shown in Table 2) do not correlate well with the ranking according to Dice and Hausdorff distance. Both FID and IS have been questioned as good metrics[16,47], but are still commonly used due to the lack of better alternatives. FID and IS focus on image quality and diversity, but do not consider memorization. Since the CNNs used for calculating FID and IS are trained on ImageNet, which only contains non-medical images, the metrics will also be biased for medical images.

Dice scores. Tables 3, 4 show the obtained Dice scores then training the segmentation networks with different combinations of real and synthetic images, and testing with real images, for BraTS 2020 and BraTS 2021

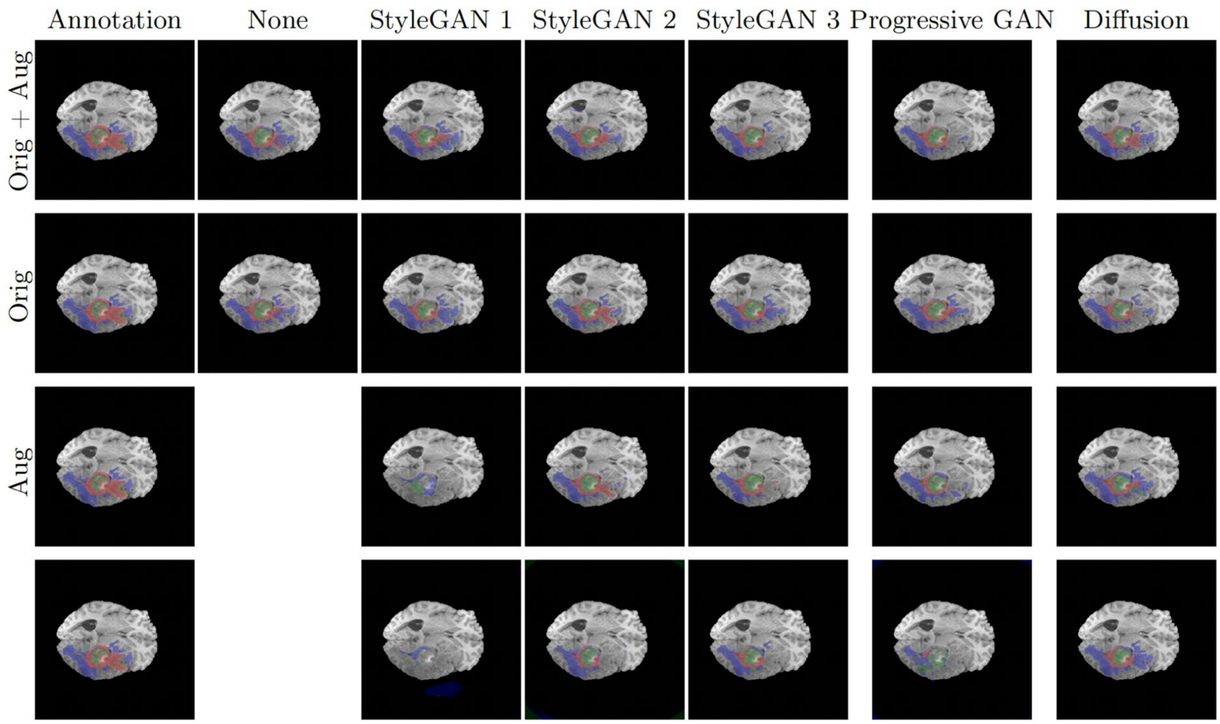

Fig. 2 Example U-Net predictions on an image in the BraTS 2020 test set. Classes are visualized as colored overlay where red is GD-enhancing tumor, blue is peritumoral edema (ED) and green is necrotic and non-enhancing tumor core (NCR/NET). Each prediction is shown for four trainings using images from each generative model; with and without augmentation and with and without the original data. The two bottom rows present predictions from when training using synthetic images.

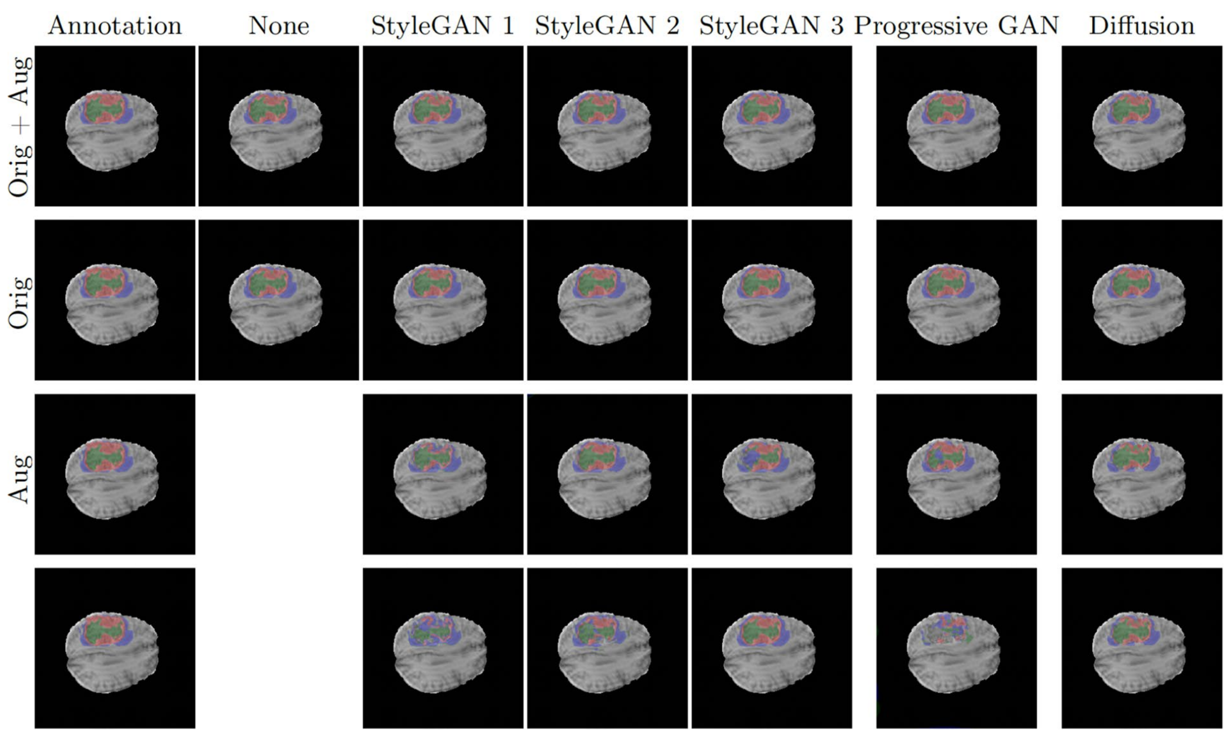

Fig. 3 Example U-Net predictions on an image in the BraTS 2021 test set. Classes are visualized as colored overlay where red is GD-enhancing tumor, blue is peritumoral edema (ED) and green is necrotic and non-enhancing tumor core (NCR/NET). Each prediction is shown for four trainings using images from each generative model; with and without augmentation and with and without the original data. The two bottom rows present predictions from when training using synthetic images.

	FID	IS	Dice Aug	Dice	Hausdorff Aug	Hausdorff
BraTS 2020			**U-Net**			
			Diffusion	Diffusion	Diffusion	Diffusion
			StyleGAN 2	StyleGAN 3	Progressive GAN	StyleGAN 3
			StyleGAN 3	StyleGAN 2	StyleGAN 2	Progressive GAN
	Diffusion	StyleGAN 3	Progressive GAN	Progressive GAN	StyleGAN 3	StyleGAN 2
	Progressive GAN	Progressive GAN	StyleGAN 1	StyleGAN 1	StyleGAN 1	StyleGAN 1
	StyleGAN 1	Diffusion	**Swin Transformer**			
	StyleGAN 3	StyleGAN 2	Diffusion	Diffusion	Diffusion	Diffusion
	StyleGAN 2	StyleGAN 1	StyleGAN 2	StyleGAN 2	StyleGAN 2	StyleGAN 3
			Progressive GAN	StyleGAN 3	StyleGAN 3	Progressive GAN
			StyleGAN 3	Progressive GAN	Progressive GAN	StyleGAN 2
			StyleGAN 1	StyleGAN 1	StyleGAN 1	StyleGAN 1
BraTS 2021			**U-Net**			
			Diffusion	Diffusion	Diffusion	Diffusion
			StyleGAN 2	StyleGAN 3	Progressive GAN	StyleGAN 3
			StyleGAN 3	StyleGAN 2	StyleGAN 3	Progressive GAN
	StyleGAN 2	StyleGAN 2	StyleGAN 1	StyleGAN 1	StyleGAN 1	StyleGAN 1
	Progressive GAN	StyleGAN 3	Progressive GAN	Progressive GAN	StyleGAN 2	StyleGAN 2
	StyleGAN 1	Diffusion	**Swin Transformer**			
	Diffusion	Progressive GAN	Diffusion	Diffusion	Diffusion	Diffusion
	StyleGAN 3	StyleGAN 1	StyleGAN 2	StyleGAN 2	StyleGAN 2	StyleGAN 3
			StyleGAN 3	StyleGAN 3	StyleGAN 3	StyleGAN 2
			Progressive GAN	Progressive GAN	Progressive GAN	StyleGAN 1
			StyleGAN 1	StyleGAN 1	StyleGAN 1	Progressive GAN

Table 1. Ranking of the five generative models based on the metrics FID, IS, Dice and Hausdorff distance (when using synthetic images only). Top rows: ranking for BraTS 2020. Bottom rows: ranking for BraTS 2021. For Dice and Hausdorff distance the models are ranked both with and without augmentation when training the segmentation network.

Model	BRATS 2020		BRATS 2021	
	FID	IS	FID	IS
Progressive GAN	18.47098	2.29209	24.58502	2.12756
StyleGAN 1	44.85624	2.02602	25.30011	2.11348
StyleGAN 2	84.77283	2.20246	20.98637	2.22750
StyleGAN 3	58.40920	2.29406	37.49911	2.19667
Diffusion model	15.85417	2.25878	32.92651	2.15857

Table 2. Comparison of the generative models using the most commonly used metrics, Fréchet inception distance (FID) and inception score (IS). A total of 100,000 synthetic T1wGd images were used to calculate each metric. While these metrics can be calculated rather quickly, they do unfortunately not correlate well with the obtained performance when training networks with synthetic images.

respectively. To make it easier to compare the performance to only using real images, Table 5 shows the relative Dice scores, i.e. the obtained mean Dice score when using real and synthetic images, or only synthetic images, divided by the obtained mean Dice score when using only real images (with augmentation).

For U-Net trained with BraTS 2020 the diffusion model results in the highest Dice scores when using only synthetic images and augmentation, followed by StyleGAN 2, StyleGAN 3 and progressive GAN. A similar ranking is obtained for the Swin transformer. Using synthetic images from StyleGAN 1 results in very low Dice scores, explained by the fact that we were not able to find good hyperparameters. When excluding StyleGAN 1, the mean Dice score when using synthetic images only is very similar for U-Net and the Swin transformer, demonstrating that the synthetic images can be used also for more recent segmentation networks. Excluding StyleGAN 1, the mean Dice score is improved by 16.8% for the U-Net when adding augmentation to synthetic images only, compared to 4.1% for the Swin transformer.

For U-Net trained with BraTS 2021, the diffusion model again results in the highest Dice scores when using only synthetic images and augmentation, followed by StyleGAN 2 and StyleGAN 3. The same ranking is obtained for the Swin transformer. Using synthetic images from StyleGAN 1 results in Dice scores that are much higher compared to for BraTS 2020, possibly explained by the fact that the hyperparameters are a better fit for this dataset. The mean Dice score when using synthetic images only is 6.3% higher for the Swin transformer compared to the U-Net, again demonstrating that the synthetic images can be used also for more recent

	Model	Orig	Aug	ET	ED	NCR/NET	Mean
U-Net	None	✓	✓	0.791 ± 0.009	0.785 ± 0.003	0.610 ± 0.008	0.729 ± 0.004
	Progressive GAN	✓	✓	0.790 ± 0.010	0.790 ± 0.007	0.609 ± 0.012	0.730 ± 0.008
	StyleGAN 1	✓	✓	0.769 ± 0.033	0.759 ± 0.047	0.593 ± 0.032	0.707 ± 0.037
	StyleGAN 2	✓	✓	**0.802 ± 0.009**	**0.794 ± 0.005**	**0.614 ± 0.009**	**0.737 ± 0.009**
	StyleGAN 3	✓	✓	0.769 ± 0.042	0.743 ± 0.060	0.582 ± 0.062	0.698 ± 0.054
	Diffusion	✓	✓	0.797 ± 0.020	0.785 ± 0.025	0.616 ± 0.014	0.733 ± 0.019
	None	✓		0.787 ± 0.010	0.783 ± 0.004	0.606 ± 0.009	0.726 ± 0.005
	Progressive GAN	✓		0.786 ± 0.010	0.784 ± 0.011	0.608 ± 0.010	0.726 ± 0.008
	StyleGAN 1	✓		0.783 ± 0.028	0.769 ± 0.035	0.602 ± 0.025	0.718 ± 0.029
	StyleGAN 2	✓		0.799 ± 0.013	0.788 ± 0.008	0.611 ± 0.011	0.732 ± 0.009
	StyleGAN 3	✓		0.777 ± 0.031	0.764 ± 0.047	0.599 ± 0.048	0.713 ± 0.041
	Diffusion	✓		0.794 ± 0.017	0.785 ± 0.017	0.613 ± 0.011	0.731 ± 0.014
	Progressive GAN		✓	0.668 ± 0.016	0.607 ± 0.048	0.468 ± 0.025	0.581 ± 0.026
	StyleGAN 1		✓	0.029 ± 0.017	0.167 ± 0.107	0.071 ± 0.036	0.089 ± 0.048
	StyleGAN 2		✓	0.750 ± 0.007	0.732 ± 0.013	0.553 ± 0.018	0.678 ± 0.009
	StyleGAN 3		✓	0.714 ± 0.021	0.709 ± 0.047	0.497 ± 0.057	0.640 ± 0.041
	Diffusion		✓	**0.791 ± 0.006**	**0.790 ± 0.005**	**0.610 ± 0.006**	**0.730 ± 0.004**
	Progressive GAN			0.546 ± 0.129	0.457 ± 0.159	0.360 ± 0.113	0.454 ± 0.130
	StyleGAN 1			0.022 ± 0.014	0.129 ± 0.099	0.057 ± 0.032	0.069 ± 0.044
	StyleGAN 2			0.562 ± 0.208	0.499 ± 0.238	0.376 ± 0.190	0.479 ± 0.206
	StyleGAN 3			0.682 ± 0.039	0.658 ± 0.067	0.444 ± 0.073	0.595 ± 0.059
	Diffusion			0.784 ± 0.009	0.783 ± 0.008	0.604 ± 0.008	0.723 ± 0.008
Swin Transformer	None	✓	✓	0.793 ± 0.015	0.803 ± 0.004	0.631 ± 0.005	0.743 ± 0.008
	Progressive GAN	✓	✓	**0.772 ± 0.005**	**0.776 ± 0.002**	**0.580 ± 0.005**	**0.709 ± 0.003**
	StyleGAN 1	✓	✓	0.754 ± 0.005	0.749 ± 0.003	0.566 ± 0.007	0.690 ± 0.005
	StyleGAN 2	✓	✓	0.770 ± 0.002	0.759 ± 0.002	0.577 ± 0.006	0.702 ± 0.006
	StyleGAN 3	✓	✓	0.755 ± 0.008	0.751 ± 0.002	0.574 ± 0.007	0.693 ± 0.006
	Diffusion	✓	✓	0.754 ± 0.009	0.755 ± 0.002	0.574 ± 0.003	0.694 ± 0.005
	None	✓		0.740 ± 0.009	0.731 ± 0.003	0.540 ± 0.005	0.670 ± 0.006
	Progressive GAN	✓		0.714 ± 0.011	0.733 ± 0.003	0.548 ± 0.005	0.665 ± 0.006
	StyleGAN 1	✓		0.732 ± 0.009	0.736 ± 0.001	0.539 ± 0.004	0.669 ± 0.005
	StyleGAN 2	✓		0.742 ± 0.012	0.742 ± 0.002	0.547 ± 0.004	0.677 ± 0.005
	StyleGAN 3	✓		0.745 ± 0.008	0.738 ± 0.002	0.562 ± 0.006	0.681 ± 0.005
	Diffusion	✓		0.749 ± 0.010	0.740 ± 0.003	0.543 ± 0.004	0.677 ± 0.006
	Progressive GAN		✓	0.667 ± 0.009	0.676 ± 0.004	0.476 ± 0.003	0.607 ± 0.003
	StyleGAN 1		✓	0.010 ± 0.002	0.149 ± 0.003	0.037 ± 0.006	0.065 ± 0.004
	StyleGAN 2		✓	0.745 ± 0.005	0.715 ± 0.003	0.426 ± 0.003	0.629 ± 0.004
	StyleGAN 3		✓	0.713 ± 0.009	0.694 ± 0.002	0.400 ± 0.003	0.603 ± 0.005
	Diffusion		✓	**0.754 ± 0.007**	**0.743 ± 0.003**	**0.556 ± 0.007**	**0.684 ± 0.006**
	Progressive GAN			0.659 ± 0.003	0.639 ± 0.005	0.458 ± 0.004	0.585 ± 0.004
	StyleGAN 1			0.024 ± 0.004	0.112 ± 0.013	0.049 ± 0.009	0.061 ± 0.008
	StyleGAN 2			0.701 ± 0.009	0.659 ± 0.002	0.404 ± 0.006	0.588 ± 0.006
	StyleGAN 3			0.701 ± 0.006	0.675 ± 0.003	0.380 ± 0.003	0.585 ± 0.004
	Diffusion			0.734 ± 0.010	0.727 ± 0.002	0.533 ± 0.004	0.664 ± 0.005

Table 3. Results on the test dataset (56 subjects) when training the generative models with BraTS 2020 (313 training subjects). Mean and standard deviation of Dice score across the dataset are calculated for the labels; GD-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET). All results are presented as mean ± standard deviation of 10 trainings. The column Orig marks if the original dataset has been added to the training set and the column Aug marks if augmentation was used during training of the segmentation network. Training with and without augmentation was performed with the same number of total images. The top row shows baseline results using only real images.

segmentation networks. The mean Dice score is improved by 15.9% for the U-Net when adding augmentation to synthetic images only, compared to only 1.8% for the Swin transformer.

Regarding relative Dice scores, Table 5 shows that that the diffusion model for U-Net trained with synthetic images only from BraTS 2020 results in the same Dice scores as when using real images, while StyleGAN 2 reaches 66%–93% and StyleGAN 3 reaches 81%–87%. For the Swin transformer, synthetic images from the diffusion model result in Dice scores that are 89%–92% compared to training with real images, while StyleGAN

	Model	Orig	Aug	ET	ED	NCR/NET	Mean
U-Net	None	✓	✓	0.847 ± 0.010	0.829 ± 0.003	0.709 ± 0.011	0.795 ± 0.006
	Progressive GAN	✓	✓	0.821 ± 0.056	0.813 ± 0.038	0.688 ± 0.035	0.774 ± 0.042
	StyleGAN 1	✓	✓	**0.853 ± 0.007**	**0.828 ± 0.004**	**0.704 ± 0.016**	**0.975 ± 0.006**
	StyleGAN 2	✓	✓	0.780 ± 0.201	0.757 ± 0.228	0.647 ± 0.200	0.728 ± 0.209
	StyleGAN 3	✓	✓	0.843 ± 0.014	0.825 ± 0.016	0.717 ± 0.020	0.795 ± 0.014
	Diffusion	✓	✓	0.835 ± 0.011	0.821 ± 0.006	0.704 ± 0.013	0.787 ± 0.007
	None	✓		0.845 ± 0.009	0.828 ± 0.003	0.708 ± 0.012	0.794 ± 0.006
	Progressive GAN	✓		0.830 ± 0.041	0.821 ± 0.028	0.692 ± 0.028	0.781 ± 0.031
	StyleGAN 1	✓		0.818 ± 0.121	0.792 ± 0.161	0.669 ± 0.151	0.760 ± 0.144
	StyleGAN 2	✓		0.810 ± 0.145	0.794 ± 0.165	0.675 ± 0.145	0.759 ± 0.151
	StyleGAN 3	✓		0.839 ± 0.015	0.826 ± 0.016	0.708 ± 0.021	0.791 ± 0.015
	Diffusion	✓		0.840 ± 0.011	0.824 ± 0.005	0.699 ± 0.014	0.788 ± 0.007
	Progressive GAN		✓	0.638 ± 0.032	0.582 ± 0.035	0.525 ± 0.026	0.582 ± 0.029
	StyleGAN 1		✓	0.710 ± 0.011	0.682 ± 0.007	0.377 ± 0.015	0.590 ± 0.007
	StyleGAN 2		✓	0.734 ± 0.022	0.738 ± 0.016	0.606 ± 0.024	0.692 ± 0.015
	StyleGAN 3		✓	0.746 ± 0.013	0.738 ± 0.004	0.473 ± 0.016	0.652 ± 0.005
	Diffusion		✓	0.757 ± 0.009	0.743 ± 0.010	0.632 ± 0.017	0.711 ± 0.009
	Progressive GAN			0.516 ± 0.142	0.400 ± 0.185	0.344 ± 0.184	0.420 ± 0.165
	StyleGAN 1			0.635 ± 0.085	0.585 ± 0.103	0.282 ± 0.098	0.501 ± 0.091
	StyleGAN 2			0.584 ± 0.156	0.571 ± 0.170	0.367 ± 0.240	0.507 ± 0.186
	StyleGAN 3			0.725 ± 0.024	0.719 ± 0.020	0.445 ± 0.031	0.629 ± 0.024
	Diffusion			**0.772 ± 0.018**	**0.764 ± 0.023**	**0.646 ± 0.020**	**0.727 ± 0.019**
Swin Transformer	None	✓	✓	0.779 ± 0.009	0.816 ± 0.005	0.674 ± 0.011	0.756 ± 0.008
	Progressive GAN	✓	✓	0.785 ± 0.012	0.812 ± 0.002	0.664 ± 0.010	0.754 ± 0.009
	StyleGAN 1	✓	✓	0.777 ± 0.008	0.817 ± 0.003	0.664 ± 0.008	0.752 ± 0.006
	StyleGAN 2	✓	✓	0.790 ± 0.006	0.813 ± 0.002	0.666 ± 0.012	0.756 ± 0.006
	StyleGAN 3	✓	✓	**0.795 ± 0.005**	**0.813 ± 0.003**	**0.671 ± 0.010**	**0.760 ± 0.006**
	Diffusion	✓	✓	**0.791 ± 0.007**	**0.816 ± 0.003**	**0.672 ± 0.009**	**0.760 ± 0.006**
	None	✓		0.772 ± 0.002	0.811 ± 0.002	0.661 ± 0.010	0.748 ± 0.005
	Progressive GAN	✓		0.781 ± 0.008	0.810 ± 0.002	0.659 ± 0.014	0.750 ± 0.008
	StyleGAN 1	✓		0.778 ± 0.008	0.810 ± 0.004	0.650 ± 0.013	0.746 ± 0.008
	StyleGAN 2	✓		0.786 ± 0.010	0.809 ± 0.003	0.656 ± 0.011	0.750 ± 0.008
	StyleGAN 3	✓		0.789 ± 0.009	0.810 ± 0.003	0.668 ± 0.015	0.756 ± 0.009
	Diffusion	✓		0.788 ± 0.010	0.814 ± 0.002	0.663 ± 0.013	0.755 ± 0.008
	Progressive GAN		✓	0.691 ± 0.009	0.672 ± 0.002	0.470 ± 0.003	0.611 ± 0.004
	StyleGAN 1		✓	0.720 ± 0.002	0.696 ± 0.002	0.375 ± 0.004	0.597 ± 0.003
	StyleGAN 2		✓	0.716 ± 0.007	0.739 ± 0.001	0.501 ± 0.003	0.652 ± 0.004
	StyleGAN 3		✓	0.731 ± 0.002	0.744 ± 0.002	0.429 ± 0.002	0.635 ± 0.002
	Diffusion		✓	**0.767 ± 0.003**	**0.794 ± 0.002**	**0.628 ± 0.010**	**0.729 ± 0.005**
	Progressive GAN			0.659 ± 0.004	0.636 ± 0.005	0.461 ± 0.004	0.585 ± 0.004
	StyleGAN 1			0.703 ± 0.007	0.683 ± 0.003	0.356 ± 0.003	0.581 ± 0.004
	StyleGAN 2			0.710 ± 0.002	0.736 ± 0.002	0.500 ± 0.002	0.649 ± 0.002
	StyleGAN 3			0.715 ± 0.009	0.732 ± 0.002	0.421 ± 0.001	0.623 ± 0.004
	Diffusion			0.768 ± 0.003	0.789 ± 0.003	0.627 ± 0.010	0.728 ± 0.005

Table 4. Results on the test dataset (56 subjects) when training the generative models with BraTS 2021 (1195 training subjects). Mean and standard deviation of Dice score across the dataset are calculated for the labels; GD-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET). All results are presented as mean ± standard deviation of 10 trainings. The column Orig marks if the original dataset has been added to the training set and the column Aug marks if augmentation was used during training of the segmentation network. Training with and without augmentation was performed with the same number of total images. The top row shows baseline results using only real images.

2 reaches 79%–84% and StyleGAN 3 reaches 78%–81%. For U-Net trained with BraTS 2021 the Dice scores obtained when training with only synthetic images are in general lower compared to BraTS 2020, except for StyleGAN 1. The diffusion model reaches 89%–91% relative Dice, while StyleGAN 2 reaches 63%–87% and StyleGAN 3 reaches 79%–82%. For the Swin transformer the relative Dice scores are in general higher compared to BraTS 2020, partly explained by the fact that the Swin transformer results in a lower Dice score than the U-Net when training with only real images (this may be explained by the fact that vision transformers normally need

	Model	Orig	Aug	Relative Dice BraTS 2020	Relative Dice BraTS 2021
U-Net	Progressive GAN	✓	✓	100.14%	97.36%
	StyleGAN 1	✓	✓	96.98%	**100.00%**
	StyleGAN 2	✓	✓	**101.10%**	91.57%
	StyleGAN 3	✓	✓	95.74%	**100.00%**
	Diffusion	✓	✓	100.55%	98.99%
	Progressive GAN	✓		99.59%	98.24%
	StyleGAN 1	✓		98.49%	95.60%
	StyleGAN 2	✓		100.41%	95.47%
	StyleGAN 3	✓		97.80%	99.5%
	Diffusion	✓		100.27%	99.12%
	Progressive GAN		✓	79.70%	73.21%
	StyleGAN 1		✓	12.21%	74.21%
	StyleGAN 2		✓	93.00%	87.04%
	StyleGAN 3		✓	87.80%	82.01%
	Diffusion		✓	**100.14%**	**89.43%**
	Progressive GAN			62.27%	52.83%
	StyleGAN 1			9.47%	63.02%
	StyleGAN 2			65.71%	63.77%
	StyleGAN 3			81.62%	79.12%
	Diffusion			99.18%	91.45%
Swin Transformer	Progressive GAN	✓	✓	**95.42%**	99.73%
	StyleGAN 1	✓	✓	92.87%	99.47%
	StyleGAN 2	✓	✓	94.48%	100.00%
	StyleGAN 3	✓	✓	93.27%	**100.52%**
	Diffusion	✓	✓	93.41%	**100.52%**
	Progressive GAN	✓		89.5%	99.20%
	StyleGAN 1	✓		90.04%	98.68%
	StyleGAN 2	✓		91.11%	99.20%
	StyleGAN 3	✓		91.65%	100.0%
	Diffusion	✓		91.11%	99.87%
	Progressive GAN		✓	81.69%	80.82%
	StyleGAN 1		✓	8.74%	78.97%
	StyleGAN 2		✓	84.66%	86.24%
	StyleGAN 3		✓	81.16%	83.99%
	Diffusion		✓	92.06%	**96.42%**
	Progressive GAN			78.73%	77.38%
	StyleGAN 1			8.21%	76.85%
	StyleGAN 2			79.13%	85.84%
	StyleGAN 3			78.73%	82.41%
	Diffusion			**89.37%**	96.30%

Table 5. Results on the test datasets (56 subjects) when training the generative models with BraTS 2020 and 2021. Relative Dice is defined as the mean Dice score when using only synthetic images, or real and synthetic, divided by the mean Dice score when using only real images with augmentation. Aug marks if augmentation was used during training of the segmentation network. Except for StyleGAN 1, the relative Dice scores are lower for BraTS 2021 when only using synthetic images. Training with and without augmentation was performed with the same number of total images.

larger datasets to perform well). The diffusion model reaches about 96% relative Dice, compared to 85%–86% for StyleGAN 2 and 82%–84% for StyleGAN 3.

To assess the impact of the ratio of real and synthetic images, we systematically increased the proportion of real images in a training set with a constant size of 100,000 images. This approach allowed us to evaluate the benefits of real data and the utility of synthetic images in enhancing model performance. The outcomes of this incremental integration are illustrated in Fig. 4, which showcases how varying the ratio of real to synthetic data affects the results. Using only 5000 real images, along with 95,000 synthetic images, still results in good performance (substantially higher compared to using 100,000 synthetic images).

Hausdorff distance. Tables 6, 7 show obtained Hausdorff distances when training the segmentation networks with synthetic images from BraTS 2020 and 2021, respectively. Overall the rankings of the generative

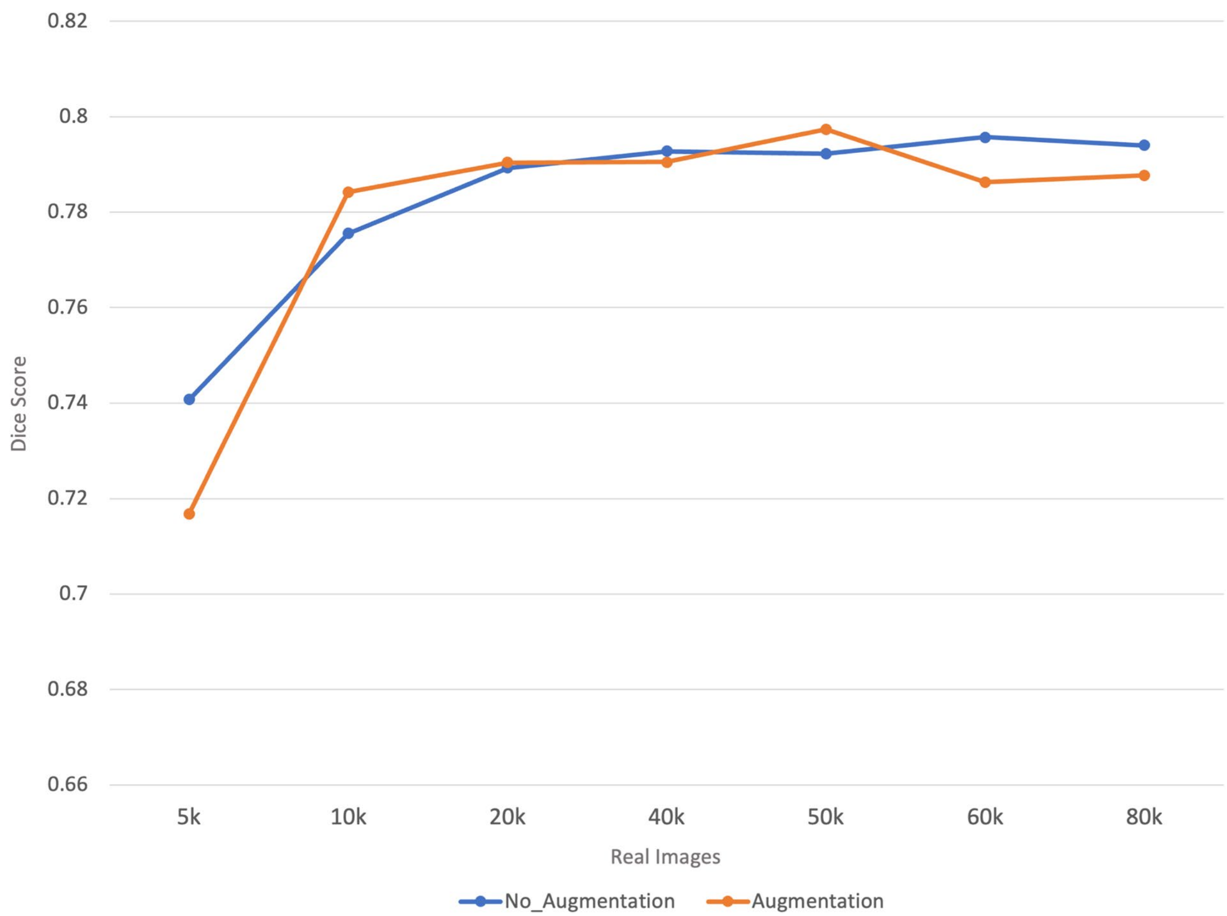

Fig. 4 Graph depicting the U-Net segmentation performance (Dice score) when using different proportions of real (BraTS 2021) and synthetic images generated from StyleGAN 3 (trained on BraTS 2021), in a constant total set of 100,000 images. As the number of real images increases along the x-axis, fewer synthetic images are used. To avoid random fluctuations, each segmentation model was trained 10 times and the average performance is presented.

models are very similar to the ranking from the Dice scores, the main difference being that the progressive GAN is ranked higher for U-Net. The mean Hausdorff distance is in general much lower (i.e. better) for the Swin transformer compared to the U-Net.

Qualitative evaluation by neuroradiologist. In addition to the quantitative metrics a qualitative evaluation by an experienced neuroradiologist was performed, see the Methods section for details. Table 8 shows the results of the evaluation, i.e. how the images were classified (real or synthetic).

Discussion

Our evaluation shows that training segmentation networks with synthetic images works well, with Dice scores that reach 91%–100% compared to when training with real images (for BraTS 2021 and 2020, respectively). Shin *et al.*[24] obtained a relative Dice score of 77.6% for brain tumor segmentation, using the deep convolutional GAN architecture which is older than progressive GAN. Fernandez *et al.*[28] obtained a similar relative Dice of 93.8%, but only performed a binary tumor segmentation which is an easier task. No comparison with other generative models was conducted. Thambawita *et al.*[27] obtained a relative Dice score of 97.2%, but for segmentation of endoscopy images making it difficult to compare the results. Furthermore, the authors did not use more recent generative models like StyleGAN or diffusion models.

The Dice scores for the diffusion model are for BraTS 2020 basically the same as when training with real images, which made us suspicious. An investigation revealed that the diffusion model had memorized many of the training images[10]. Memorization has previously been shown when using diffusion models for non-medical images[8,9], but to the best of our knowledge not for medical images. Diffusion models are more likely to memorize the training images compared to GANs[8,10], due to a completely different architecture.

Even better results can be obtained using an ensemble of 5–10 generative models[16,22,48], as each model by random chance will learn a different subset of the high dimensional distribution, at the cost of a training time which is 5–10 times longer. Larsson *et al.*[22] demonstrated that using an ensemble of 10 progressive GANs improved the mean Dice score for brain tumor segmentation by 9.5%, compared to a single GAN. The benefit of using an

	Model	Orig	Aug	ET	ED	NCR/NET	Mean
U-Net	None	✓	✓	16.4 ± 1.8	33.8 ± 3.4	19.2 ± 1.2	23.1 ± 1.9
	Progressive GAN	✓	✓	15.9 ± 1.6	33.8 ± 2.9	21.5 ± 1.9	23.7 ± 1.8
	StyleGAN 1	✓	✓	26.6 ± 16.0	42.6 ± 11.1	24.4 ± 4.0	31.2 ± 9.9
	StyleGAN 2	✓	✓	**15.4 ± 1.9**	**32.7 ± 2.6**	**19.6 ± 1.4**	**22.6 ± 1.5**
	StyleGAN 3	✓	✓	31.3 ± 22.7	44.4 ± 15.4	33.7 ± 19.2	36.5 ± 18.7
	Diffusion	✓	✓	21.7 ± 20.1	32.3 ± 8.4	22.3 ± 14.0	25.4 ± 14.1
	None	✓		17.7 ± 2.4	35.9 ± 4.5	21.3 ± 2.7	24.9 ± 2.8
	Progressive GAN	✓		17.7 ± 3.1	35.2 ± 4.2	22.6 ± 2.0	25.1 ± 2.5
	StyleGAN 1	✓		22.3 ± 12.4	39.1 ± 9.1	23.5 ± 3.2	28.3 ± 7.8
	StyleGAN 2	✓		16.3 ± 3.2	34.2 ± 2.8	20.2 ± 1.8	23.6 ± 2.0
	StyleGAN 3	✓		24.5 ± 17.6	39.2 ± 12.3	26.9 ± 15.3	30.2 ± 14.7
	Diffusion	✓		19.1 ± 14.5	32.9 ± 6.5	21.4 ± 10.0	24.5 ± 10.1
	Progressive GAN		✓	35.2 ± 6.7	44.9 ± 4.3	34.0 ± 4.3	38.0 ± 3.7
	StyleGAN 1		✓	82.1 ± 6.0	90.1 ± 40.5	108.6 ± 49.4	93.6 ± 29.6
	StyleGAN 2		✓	33.4 ± 7.9	62.4 ± 24.5	36.4 ± 11.4	44.1 ± 11.9
	StyleGAN 3		✓	37.8 ± 12.2	49.6 ± 7.5	50.4 ± 7.8	45.9 ± 8.9
	Diffusion		✓	**15.2 ± 7.1**	**32.1 ± 3.9**	**26.3 ± 10.6**	**24.6 ± 5.8**
	Progressive GAN			58.6 ± 40.9	94.6 ± 58.3	70.1 ± 46.7	74.4 ± 44.9
	StyleGAN 1			103.7 ± 28.7	112.2 ± 39.1	131.2 ± 44.5	115.7 ± 33.5
	StyleGAN 2			106.6 ± 78.2	116.8 ± 61.7	97.7 ± 73.6	107.0 ± 68.0
	StyleGAN 3			48.8 ± 14.6	58.9 ± 10.9	64.0 ± 17.7	57.2 ± 13.6
	Diffusion			16.7 ± 5.7	34.9 ± 5.4	28.4 ± 9.5	26.6 ± 5.2
Swin Transformer	None	✓	✓	16.1 ± 1.2	27.5 ± 1.7	19.3 ± 1.4	20.3 ± 0.8
	Progressive GAN	✓	✓	25.9 ± 2.1	47.9 ± 1.6	47.6 ± 2.3	38.3 ± 1.7
	StyleGAN 1	✓	✓	13.8 ± 1.1	26.1 ± 0.9	18.2 ± 1.1	18.7 ± 1.0
	StyleGAN 2	✓	✓	14.5 ± 1.1	27.1 ± 0.8	19.0 ± 1.3	19.8 ± 0.9
	StyleGAN 3	✓	✓	14.5 ± 0.7	25.2 ± 0.7	19.4 ± 2.4	19.3 ± 1.0
	Diffusion	✓	✓	**14.5 ± 0.9**	**25.7 ± 1.0**	**19.4 ± 1.1**	**19.2 ± 0.7**
	None	✓		16.8 ± 0.6	29.6 ± 1.0	20.0 ± 1.0	21.4 ± 0.5
	Progressive GAN	✓		16.8 ± 0.7	30.4 ± 1.3	20.3 ± 0.9	21.6 ± 0.8
	StyleGAN 1	✓		16.6 ± 1.3	28.3 ± 1.1	19.6 ± 0.8	20.6 ± 0.6
	StyleGAN 2	✓		15.7 ± 0.9	28.8 ± 1.7	19.4 ± 0.9	20.4 ± 0.9
	StyleGAN 3	✓		15.3 ± 1.3	26.8 ± 1.4	17.9 ± 1.7	19.4 ± 1.2
	Diffusion	✓		16.0 ± 0.6	28.0 ± 1.3	19.7 ± 0.7	20.9 ± 0.5
	Progressive GAN		✓	26.6 ± 1.7	48.5 ± 1.0	47.4 ± 1.6	38.4 ± 0.9
	StyleGAN 1		✓	75.6 ± 4.2	72.3 ± 2.1	87.9 ± 2.1	74.8 ± 3.1
	StyleGAN 2		✓	19.2 ± 1.5	37.8 ± 2.0	47.7 ± 2.8	33.5 ± 1.7
	StyleGAN 3		✓	20.5 ± 1.0	38.8 ± 1.5	48.0 ± 1.6	34.0 ± 1.4
	Diffusion		✓	**14.7 ± 0.8**	**25.2 ± 1.0**	**20.0 ± 1.2**	**19.5 ± 0.7**
	Progressive GAN			38.7 ± 1.9	63.3 ± 1.1	69.9 ± 1.9	54.6 ± 0.4
	StyleGAN 1			74.6 ± 2.8	66.5 ± 2.7	82.7 ± 2.7	70.0 ± 2.3
	StyleGAN 2			47.4 ± 3.1	64.0 ± 0.5	79.1 ± 0.9	60.1 ± 1.2
	StyleGAN 3			23.4 ± 2.7	39.6 ± 2.2	51.6 ± 1.7	36.7 ± 1.9
	Diffusion			16.2 ± 0.9	27.5 ± 1.3	20.3 ± 1.3	20.8 ± 0.9

Table 6. Results on the test dataset (56 subjects) when training the generative models with BraTS 2020 (313 training subjects). Mean and standard deviation of Hausdorff distance across the dataset are calculated for the labels; GD-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET). All results are presented as mean ± standard deviation of 10 trainings. The column Orig marks if the original dataset has been added to the training set and the column Aug marks if augmentation was used during training of the segmentation network. Training with and without augmentation was performed with the same number of total images. The top row shows baseline results using only real images. The bottom 10 rows show results when training with only synthetic images, with and without augmentation.

ensemble is expected to be larger for BraTS 2021, compared to BraTS 2020 used in[22], to capture all modes of the distribution (due to a larger number of imaging sites in BraTS 2021).

Training the generative models with 1195 subjects (BraTS 2021), instead of 313 (BraTS 2020), leads to worse performance for the U-Net (lower relative Dice scores when using only synthetic images, mean 83.68% versus 77.36%, excluding StyleGAN 1) which may seem surprising. However, BraTS 2021 contains data from a larger

	Model	Orig	Aug	ET	ED	NCR/NET	Mean
U-Net	None	✓	✓	12.3 ± 2.0	22.2 ± 2.1	12.1 ± 1.7	15.5 ± 1.6
	Progressive GAN	✓	✓	30.5 ± 49.1	30.6 ± 14.9	12.9 ± 2.7	24.7 ± 21.2
	StyleGAN 1	✓	✓	**13.3 ± 0.4**	**23.5 ± 3.4**	**12.0 ± 1.2**	**16.2 ± 1.4**
	StyleGAN 2	✓	✓	25.7 ± 36.2	37.4 ± 47.0	25.4 ± 42.7	29.5 ± 41.9
	StyleGAN 3	✓	✓	13.8 ± 4.4	24.8 ± 7.9	12.8 ± 1.0	17.1 ± 4.2
	Diffusion	✓	✓	16.2 ± 3.3	26.6 ± 4.5	12.4 ± 1.2	18.4 ± 1.6
	None	✓		12.8 ± 1.7	22.5 ± 2.2	11.9 ± 1.3	15.7 ± 1.3
	Progressive GAN	✓		21.6 ± 35.9	28.4 ± 10.9	12.3 ± 2.1	20.8 ± 15.5
	StyleGAN 1	✓		18.9 ± 21.6	29.1 ± 24.2	18.8 ± 30.0	22.3 ± 25.2
	StyleGAN 2	✓		19.2 ± 26.5	30.4 ± 34.0	18.5 ± 30.9	22.7 ± 30.4
	StyleGAN 3	✓		14.7 ± 5.4	26.4 ± 7.8	12.6 ± 1.6	17.9 ± 4.6
	Diffusion	✓		14.9 ± 3.1	25.0 ± 3.9	12.3 ± 0.9	17.4 ± 1.8
	Progressive GAN		✓	31.2 ± 5.3	48.0 ± 4.4	23.9 ± 1.6	34.3 ± 3.0
	StyleGAN 1		✓	52.8 ± 14.2	76.7 ± 31.1	99.8 ± 21.3	76.4 ± 15.9
	StyleGAN 2		✓	78.5 ± 53.3	124.6 ± 46.2	115.8 ± 57.6	106.3 ± 41.5
	StyleGAN 3		✓	30.8 ± 7.3	50.5 ± 3.1	63.6 ± 3.5	48.3 ± 3.2
	Diffusion		✓	28.8 ± 8.8	41.9 ± 4.6	15.1 ± 4.1	28.6 ± 3.9
	Progressive GAN			86.9 ± 64.8	112.8 ± 65.1	108.6 ± 84.9	102.7 ± 69.6
	StyleGAN 1			108.6 ± 57.3	124.6 ± 53.1	142.8 ± 46.9	125.3 ± 50.7
	StyleGAN 2			106.0 ± 48.1	142.4 ± 40.6	151.3 ± 55.7	133.2 ± 41.2
	StyleGAN 3			31.0 ± 5.7	52.2 ± 3.6	67.0 ± 4.8	50.0 ± 3.6
	Diffusion			**31.7 ± 14.7**	**38.0 ± 5.9**	**15.5 ± 3.7**	**28.4 ± 5.4**
Swin Transformer	None	✓	✓	11.3 ± 0.4	17.2 ± 0.7	11.4 ± 0.8	11.5 ± 0.3
	Progressive GAN	✓	✓	11.5 ± 0.9	18.6 ± 0.9	11.2 ± 0.9	11.7 ± 0.6
	StyleGAN 1	✓	✓	**10.4 ± 0.8**	**18.1 ± 0.9**	**10.9 ± 0.5**	**11.1 ± 0.3**
	StyleGAN 2	✓	✓	11.1 ± 0.7	17.2 ± 1.1	11.1 ± 0.5	11.2 ± 0.5
	StyleGAN 3	✓	✓	11.2 ± 0.9	17.5 ± 1.1	11.1 ± 0.5	11.3 ± 0.6
	Diffusion	✓	✓	11.0 ± 0.5	16.9 ± 0.6	11.5 ± 0.6	11.2 ± 0.5
	None	✓		11.4 ± 0.9	18.1 ± 1.0	11.4 ± 0.9	11.6 ± 0.8
	Progressive GAN	✓		11.3 ± 0.6	18.9 ± 0.8	11.1 ± 0.6	11.7 ± 0.7
	StyleGAN 1	✓		11.1 ± 0.7	18.1 ± 1.2	11.1 ± 0.3	11.2 ± 0.6
	StyleGAN 2	✓		11.6 ± 0.9	17.3 ± 1.3	11.2 ± 0.5	11.2 ± 0.9
	StyleGAN 3	✓		11.2 ± 0.3	18.1 ± 0.9	11.2 ± 0.3	11.8 ± 0.4
	Diffusion	✓		11.5 ± 0.7	17.3 ± 1.1	11.5 ± 0.7	11.6 ± 0.3
	Progressive GAN		✓	26.3 ± 1.8	49.9 ± 1.3	55.0 ± 1.4	40.2 ± 1.0
	StyleGAN 1		✓	18.1 ± 1.5	41.1 ± 1.9	62.5 ± 1.4	37.1 ± 1.10
	StyleGAN 2		✓	14.0 ± 0.8	31.6 ± 1.7	42.5 ± 2.1	25.7 ± 1.0
	StyleGAN 3		✓	16.0 ± 1.3	38.4 ± 2.6	44.5 ± 1.2	29.2 ± 1.3
	Diffusion		✓	**12.7 ± 0.7**	**17.7 ± 0.6**	**14.1 ± 0.5**	**12.7 ± 0.6**
	Progressive GAN			49.2 ± 1.9	63.4 ± 0.7	72.8 ± 0.8	58.2 ± 0.8
	StyleGAN 1			24.3 ± 1.8	58.9 ± 3.1	73.9 ± 0.7	49.1 ± 1.2
	StyleGAN 2			19.3 ± 1.2	47.0 ± 1.6	58.9 ± 1.5	38.4 ± 1.1
	StyleGAN 3			16.8 ± 1.6	43.3 ± 1.9	49.6 ± 2.2	32.5 ± 1.4
	Diffusion			13.7 ± 0.9	19.2 ± 0.8	13.3 ± 0.7	13.0 ± 0.8

Table 7. Results on the test dataset (56 subjects) when training the generative models with BraTS 2021 (1195 training subjects). Mean and standard deviation of Hausdorff distance across the dataset are calculated for the labels; GD-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET). All results are presented as mean ± standard deviation of 10 trainings. The column Orig marks if the original dataset has been added to the training set and the column Aug marks if augmentation was used during training of the segmentation network. Training with and without augmentation was performed with the same number of total images. The top row shows baseline results using only real images. The bottom 10 rows show results when training with only synthetic images, with and without augmentation.

number of sites (23 versus 19), which will result in more modes in the high dimensional distribution, which is harder to learn. Furthermore, using a larger dataset like BraTS 2021 makes it harder for the generative models to memorize the training images[9,10]. It would be very interesting to compare our results to SinGAN-Seg[27], where a single image is used to train the generative model, but we suspect that such a model is prone to memorization.

Image type	Classified as real	Classified as synthetic
Real image	**16**	84
Progressive GAN	23	**77**
StyleGAN 1	14	**86**
StyleGAN 2	13	**87**
StyleGAN 3	18	**82**
Diffusion Model	16	**84**

Table 8. Results from the qualitative evaluation by an experienced neuroradiologist, who classified 600 four-panel images as real or synthetic (100 real images and 100 images per generative model). The two right columns show how many images that were classified as real or synthetic, for real images and for synthetic images from the five generative models. The number of correct classifications for each row has been marked with bold.

Our results show that augmentation makes a rather big difference for the U-Net when training with only synthetic images, while the improvement is smaller for the Swin transformer. A possible explanation for this is that the augmentation helps the segmentation networks to overcome systematic differences between real and synthetic images. To apply augmentation when training the generative models needs to be explored in future work, as it on the one hand can increase the number of training images, but on the other hand it may introduce more modes in the high dimensional distribution (which will be harder to learn).

Several other researchers have demonstrated that combining real and synthetic images (i.e. using generative models for advanced augmentation) can improve segmentation accuracy[20,21,24], or classification accuracy[13,18], compared to training with only real images, but our results show that adding synthetic images only provides minor improvements or even results in worse performance. There are at least two possible explanations for this. First, we use a rather strong baseline segmentation model with several types of traditional augmentation during training. Second, we repeat the training of each segmentation network 10 times to avoid differences due to random chance. It is possible that using a subset of the real data (e.g. 20%), instead of all the real data, would result in larger improvements when adding synthetic images. However, the generative models should then be trained with the same subset, which will reduce the quality and diversity of the synthetic images.

The results are likely to strongly depend on the hyperparameters and the architecture of each generative model, but to explore many parameter combinations and architectures is difficult due to the long training times of both generative models and segmentation networks. For this reason, the results presented in this work may not correspond to the optimal results for each generative model, which one could obtain after performing an exhaustive optimization of hyperparameters and architectures. The total training time for this work was over 2000 days on an Nvidia A100 graphics card, and an exhaustive search of hyperparameters could have increased this a factor 10. This work demonstrates that new efficient metrics for evaluating synthetic medical images are required, as FID and IS are based on ImageNet (which does not contain medical images), do not consider memorization[8–10], and do in general not correlate with how a network trained on the synthetic images will perform (see Table 1)[16,47]. In future work we will calculate FID and IS using CNNs pre-trained on RadImageNet[49], which is a large collection of medical images, to see if Rad-FID and Rad-IS better correlate with our other metrics.

Regarding the qualitative evaluation by a neuroradiologist, the results show that the generative models produce synthetic images that are on the same level as real images (a similar number of images were classified as synthetic). It should however be noted that the setup of this experiment was not similar to a regular clinical assessment of brain tumor MRI, which is reflected in the fact that a large portion of the real images were falsely classified as synthetic images. Normally, in a clinical workflow a neuroradiologist would assess the whole brain, instead of a single slice, with even more MRI sequences than used in this study. Furthermore, a neuroradiologist does not normally look at skullstripped brain images. A challenge with this evaluation was that the BraTS images originate from many different scanners, with varying image quality, which probably also affected the visual assessment. Two additional limitations are that only one neuroradiologist performed the visual assessment, and that the sample of 600 images is not balanced in terms of real and synthetic images (which may introduce a bias).

The implication of our results are that sharing synthetic medical images is a viable option to sharing real images. A researcher can use synthetic images for pre-training, and then fine-tune the model on a small number of locally available images. Sharing synthetic medical images can be substantially easier[11,12,27,50], as GDPR should not apply for data which do not belong to a specific person (but further legal research is needed). Regarding consent, Larson *et al.*[51] argue that clinical data should be treated as a form of public good, to be used for the benefit of future patients, and further argue that consent is not required before collected data are used for secondary purposes when obtaining such consent is prohibitively costly or burdensome (e.g. contacting 1,000–10,000 persons). On the other hand, the argument of clinical data being treated as a form of public good, and not requiring further consent for use in research or development, may be a slippery slope and many examples exist where a retrospective look identifies the continued use of such data as an unauthorized abuse (e.g. Henrietta Lacks).

Before sharing synthetic images it is important to investigate how similar each synthetic image is to all training images, as especially diffusion models have been shown to memorize the training images[8–10,52]. This is extra important for small datasets, as memorization is then more likely[9,10]. Common evaluation metrics like FID and IS do not capture memorization, and it is therefore necessary to for example calculate the correlation, or some other metric like mutual information, between each synthetic image and all training images[10,48,52]. Pre-trained generative models can play an important role for sharing synthetic images from small datasets, as it should be less likely for a pre-trained model to memorize a small number of new images during fine tuning (compared to

Model	Batch size	Iterations	gamma	Hardware	Training time
Progressive GAN	128 - 8	12000	—	1 × Nvidia V100	5 days
StyleGAN 1	16	25000	10	2 × Nvidia V100	6 days
StyleGAN 2	32	25000	8.2	4 × Nvidia A100	3 days
StyleGAN 3	32	25000	8.2	4 × Nvidia A100	3 days
Diffusion model	12	7800	—	1 × Nvidia V100	22 days

Table 9. The used hyperparameters for the different generative models, as well as the used hardware and the approximate training time. One iteration corresponds to 1000 images processed. Processing 1000 images for the diffusion model is much slower, since it performs 4000 diffusion steps for each batch.

training the model from scratch). To determine an acceptable range of overlap with real clinical data is a very difficult task, especially since different legal experts interpret GDPR differently (in Sweden it is in general interpreted stricter compared to other countries) and since this acceptable range is likely to be different for different types of medical images. It therefore remains an open question how high the highest similarity can be before a synthetic image is seen as a copy of a training image.

Methods

Data. The MR images used for this project were downloaded from the Multimodal Brain tumour Segmentation Challenge (BraTS) 2020 and 2021[36–41]. The training set contains MR volumes of shape 240 × 240 × 155 from 369 subjects for BraTS 2020 and from 1251 subjects for BraTS 2021. For each subject four types of MR images are available: T1-weighted (T1w), post gadolinium contrast T1-weighted (T1wGd), T2-weighted (T2w), and T2-weighted fluid attenuated inversion recovery (FLAIR). The annotations cover three parts of the brain tumor: peritumoural edema (ED), necrotic and non-enhancing tumour core (NCR/NET), and GD-enhancing tumour (ET). We used 313/1195 subjects for training and 56 subjects for testing, after first performing a random shuffling of the subjects. The data in the test sets were not used for training the generative models.

All 3D volumes were split into 2D slices, as a 2D GAN and a 2D diffusion model were used (3D GANs and 3D diffusion models are not yet very common). Only slices with at least 15% pixels with an intensity of more than 50 were included in the training. This resulted in a total of 23,478 5-channel images for BraTS 2020, and 91,271 5-channel images for BraTS 2021. Each slice was zero padded from 240 × 240 to 256 × 256 pixels, as the used GANs only work for resolutions that are a power of 2, and the intensity was rescaled to 0–255. The intensities for the tumor annotations were changed from[1,2,4] to [51,102,204], such that the intensity range is more similar for the 5 channels.

Image generation. In this work we compare four different GANs (progressive growing GAN[30], StyleGAN 1[31], StyleGAN 2[32], StyleGAN 3[33]) and a diffusion model[34,35], for the task of generating brain tumor images. GANs are trained through adversarial learning (using an adversarial loss function), where a generator and a discriminator compete against each other, to produce more realistic images and to be better at discriminating images as real or synthetic. At inference time, only the generator is used. A diffusion model, on the other hand, starts with real data samples, progressively adds noise over many steps according to a predetermined schedule until the data becomes pure noise. In this work a linear noise scheduler was used. The diffusion model is then trained to reverse this process, using more traditional loss functions, reconstructing less noisy data from more noisy data at each step. During inference, a diffusion model starts with an image of pure noise and sequentially applies the learned denoiser to reduce the noise, following the reverse of the training noise schedule. This process is iterated until the noise is completely removed, resulting in the generation of a new image which resembles the training data distribution. In general diffusion models are easier to train compared to GANs, due to more traditional loss functions, but are much slower at generating images.

The openly available code of each generative model was modified to generate 5-channel images instead of 3 channels, no other modifications to the default architectures were done. Each generative model will thereby learn to jointly generate the four MR images (T1w, T1wGD, T2w, FLAIR) and the corresponding tumor annotation at the same time. There is no guarantee that the synthetic annotations will be restricted to the same values as the real annotations ([51,102,204]). The synthetic annotations were therefore thresholded to the closest original annotation value.

The used hyperparameters of each model, and the approximate training times, are provided in Table 9. We used a set of common hyperparameters across all models, along with some model-specific ones. For instance, in the case of StyleGANs, we experimented with different gamma values, the best of which are detailed in the accompanying table. For the diffusion models, they were trained with varying diffusion steps, but the optimal results were obtained with 4000 steps for both training and inference (decided by visual inspection).

For each generative model a total of 100,000 synthetic 5-channel images were generated. For the GANs this took about 10 minutes, while it took 1.5 days (using 8 GPUs) for the diffusion model The synthetic images and the trained generative models are shared on AIDA data hub[44,45].

Quantitative evaluation and tumor segmentation. The quality and diversity of synthetic images are often evaluated using metrics such as Frechét inception distance (FID) and inception score (IS)[46]. Since these metrics are based on CNNs trained on ImageNet, which does not contain medical images, they will be biased for medical images. Furthermore, these metrics will not tell us how well a network trained with synthetic images will perform on real images. The synthetic images were therefore used to train segmentation networks (based on U-Net and Swin transformers), and the evaluation was performed using real images in the test set. To investigate how FID and IS correlate with the performance of training with only synthetic images, FID and IS were also calculated.

Training deep networks is a stochastic process, meaning that training the same model several times will give different results. Each segmentation network was therefore trained 10 times, to make sure that performance differences between the different generative models are not due to random chance. The segmentation was performed for each slice independently, and the Dice scores and Hausdorff distances were then calculated in 3D after putting the slices for each subject back into a volume.

U-Net. The model structure and training setup used was inspired by the 2D segmentation code from nnUNet[53]. The model was a U-Net[54] with an extra depth layer and instance normalization instead of batch normalization. In addition, the ReLU activations was swapped for leaky ReLUs with a negative slope of 10^{-2}. All models were trained with a loss consisting of a cross-entropy term and a soft Dice term weighted equally. In addition, deep supervision was used, meaning that the loss was applied on the five highest depth level with weighting 0.5^d where d is the depth.

The loss was minimized using stochastic gradient descent with Nesterov momentum of 0.99 and weight decay of $3{\cdot}10^{-5}$. The initial learning rate was $5{\cdot}10^{-2}$ and was decreased using polynomial learning rate decay with an exponent of 0.9. The learning rate and optimizer was different for the generative models, see their respective paper. All models were trained for $3{\cdot}10^7$ samples or 3 days, whichever occured first. 20% of the available images were used for validation and the model with the best mean Dice over the validation set was used for evaluation. If both real and synthetic data was used during training, the real dataset and the synthetic dataset were sampled equally often.

During training a geometric, and intensity-augmentation was applied, as our previous work on augmentation for brain tumor segmentation[55] demonstrated that augmentation can provide significant improvements even if the dataset is large. The image and target is first randomly rotated and scaled. Both rotation and scaling is applied with a probability of 0.75, the image and target is rotated with an angle uniformly sampled from [−30°, 30°] and the width and height is scaled (independently) with a scale factor uniformly sampled from [0.9, 1.1]. Then the four input channels are augmented by; adding Gaussian noise (applied with probability 0.5, zero mean and standard deviation uniformly sampled from [0.0, 0.05]), blurring the image (applied with probability 0.2, Gaussian blurring with standard deviation sampled from [0.5, 1.0]), faking lower resolution (zooming with a factor between 0.75 and 1.0 and then upsampling) and changing the gamma factor (scaling it with a factor between 0.8 and 1.2). Lastly the input image channels are normalized using Z-score normalization.

Swin transformers. The Swin transformer segmentation network was implemented and trained using the MMSegmentation library[56]. The architecture employed is the Swin-Base variant, as implemented in MMSegmentation, with a window size of 7 and a patch size of 4×4. The original Swin transformer was designed for 3-channel RGB images, hence, to accommodate MRI scans with four modalities per slice, the number of input channels in the model was modified to 4. Consequently, the input dimension of the Swin transformer is set to $256 \times 256 \times 4$, where the '4' denotes the number of channels, and '256' represents both the height and width of each modality slice. The sole alteration to the original Swin-Base architecture is the adaptation of both the number of channels and the number of classes to 4.

The network was trained using DiceCELoss function. Furthermore, the Dice Loss component does not take into account the background class, label 0. The loss per batch was derived by calculating the loss for each training image, and taking the mean loss value.

The AdamW optimizer with $\beta\_1 = 0.9$, $\beta\_2 = 0.999$ and weight decay $\lambda = 0.01$ was used. Additionally, a learning rate scheduler with warmup and linear decay was employed. The warmup ratio was set to $1e^{-6}$. For the first 1500 iterations, the learning rate is increased and after 1500 warmup iterations, the learning rate reached 0.00006. For the rest of the training, the learning rate was decreased linearly until it reached 0.0 at the end of the training. The models were trained for 25 epochs, and after every epoch, a validation loss was calculated. The batch size was set to 8, and the last batch is dropped if it is not the same size as all of the other batches to ensure that all models are provided with batches of consistent size.

Data augmentation was performed using five techniques inspired by Cirillo *et al.*[55]. Images and their corresponding segmentations undergo rotation, with angles from 0° to 30° chosen randomly, and scaling, with axis factors varying within ±20%. Images were also subjected to a 50% chance of either horizontal or vertical flipping. Brightness was adjusted through a power-law γ intensity transformation, with parameters randomly picked between 0.8 and 1.2. Lastly, elastic deformation, following the methodology from the original U-Net paper[42], was applied using a deformation grid with normal distribution displacements ($\sigma = 2$ voxels) and smoothed with a third-order spline filter.

Qualitative evaluation by neuroradiologist. To evaluate how the synthetic images are perceived by a clinician, a total of 600 4-sequence-panel images (T1w, T1wGd, T2w, T2w FLAIR) were presented to an experienced (13 years) neuroradiologist (co-author IB). The task was to determine if each presented 4-panel image was real or synthetic. The 600 images consisted of 100 real images, and 500 synthetic images, 100 each from the five generative models (progressive GAN, StyleGAN 1–3, diffusion model). The total number of real images, and the number of synthetic images per generative model, was known to the neuroradiologist before starting the evaluation. The real and synthetic images were presented in a random order. The evaluation took approximately 12 hours. Figure 5 shows an example of a real and a synthetic image presented during the evaluation.

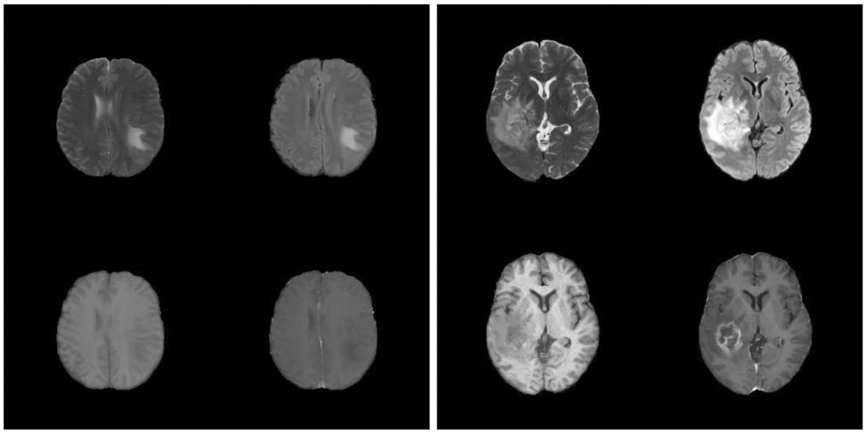

Fig. 5 Left: a real 4-channel image shown during the qualitative evaluation, where the task was to classify each example as real or synthetic. Right: a synthetic 4-channel image shown during the qualitative evaluation.

Ethics. This research study was conducted retrospectively using anonymized human subject data made available by BraTS. The ethical review board of Linköping decided that no further ethical approval was required.

Data availability

The BraTS 2020 and 2021 datasets are openly available through the following websites.

https://www.med.upenn.edu/cbica/brats2020/data.html
http://braintumorsegmentation.org

The generated synthetic images (100,000 five channel images per generative model), and the trained generative models, are shared at the AIDA data hub[44,45]; https://datahub.aida.scilifelab.se/10.23698/aida/synthetic/brgandi.

Code availability

The code for the different GANs and the diffusion model is openly shared by the creators of the generative models, see below. We therefore share our modifications to make the code work for 5-channel images, instead of 3-channel images, the used segmentation code, and some additional help scripts.
https://github.com/muhamadusman/Assist/
Generative models
Progressive growing GAN, https://github.com/tkarras/progressive_growing_of_gans
StyleGAN 1, https://github.com/NVlabs/stylegan
StyleGAN 2, https://github.com/NVlabs/stylegan2
StyleGAN 3, https://github.com/NVlabs/stylegan3
Diffusion model, https://github.com/openai/guided-diffusion
Segmentation models
U-Net, https://github.com/MIC-DKFZ/nnUNet
Swin Transformer, https://github.com/open-mmlab/mmsegmentation.

Received: 6 July 2023; Accepted: 15 February 2024;
Published online: 29 February 2024

References

1. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017).
2. Jiang, F. *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* **2** (2017).
3. Wong, J. *et al.* Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiotherapy and Oncology* **144**, 152–158 (2020).
4. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
5. Van Essen, D. C. *et al.* The wu-minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013).
6. Markiewicz, C. J. *et al.* The openneuro resource for sharing of neuroscience data. *eLife* **10**, e71774 (2021).
7. Littlejohns, T. J. *et al.* The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications* **11**, 2624 (2020).
8. Carlini, N. *et al.* Extracting training data from diffusion models. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5253-5270) (2023).
9. Somepalli, G., Singla, V., Goldblum, M., Geiping, J. & Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6048–6058 (2023).
10. Akbar, M. U., Wang, W. & Eklund, A. Beware of diffusion models for synthesizing medical images - a comparison with GANs in terms of memorizing brain MRI and chest x-ray images. *arXiv:2305.07644* (2023).
11. Rankin, D. *et al.* Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics* **8**, e18910 (2020).
12. El Emam, K., Mosquera, L., Jonker, E. & Sood, H. Evaluating the utility of synthetic COVID-19 case data. *JAMIA open* **4**, ooab012 (2021).

13. Frid-Adar, M. *et al*. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).
14. Guan, S. & Loew, M. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. *Journal of Medical Imaging* **6**, 031411–031411 (2019).
15. Qin, Z., Liu, Z., Zhu, P. & Xue, Y. A gan-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine* **195**, 105568 (2020).
16. Eilertsen, G., Tsirikoglou, A., Lundström, C. & Unger, J. Ensembles of GANs for synthetic training data generation. *ICLR 2021 workshop on Synthetic Data Generation* (2021).
17. Coyner, A. S. *et al*. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity diagnosis. *Ophthalmology Science* **2**, 100126 (2022).
18. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M. & Fleet, D. J. Synthetic data from diffusion models improves ImageNet classification. *arXiv:2304.08466* (2023).
19. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: A review. *Medical image analysis* **58**, 101552 (2019).
20. Bowles, C. *et al*. GAN augmentation: Augmenting training data using generative adversarial networks. *arXiv:1810.10863* (2018).
21. Pollastri, F., Bolelli, F., Paredes, R. & Grana, C. Augmenting data with GANs to segment melanoma skin lesions. *Multimedia Tools and Applications* **79**, 15575–15592 (2020).
22. Larsson, M., Akbar, M. U. & Eklund, A. Does an ensemble of GANs lead to better performance when training segmentation networks with synthetic images? *arXiv:2211.04086* (2022).
23. Guibas, J. T., Virdi, T. S. & Li, P. S. Synthetic medical images from dual generative adversarial networks. *arXiv:1709.01872* (2017).
24. Shin, H.-C. *et al*. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018*, *Proceedings 3*, 1–11 (Springer, 2018).
25. Foroozandeh, M. & Eklund, A. Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE. *arXiv:2009.05946* (2020).
26. Shao, S. *et al*. DiffuseExpand: Expanding dataset for 2D medical image segmentation using diffusion models. *arXiv:2304.13416* (2023).
27. Thambawita, V. *et al*. SinGAN-Seg: Synthetic training data generation for medical image segmentation. *PloS one* **17**, e0267976 (2022).
28. Fernandez, V. *et al*. Can segmentation models be trained with fully synthetically generated data? In *Simulation and Synthesis in Medical Imaging: 7th International Workshop, SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022*, *Proceedings*, 79–90 (Springer, 2022).
29. Park, T., Liu, M.-Y., Wang, T.-C. & Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346 (2019).
30. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *ICLR* (2018).
31. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410 (2019).
32. Karras, T. *et al*. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119 (2020).
33. Karras, T. *et al*. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* **34**, 852–863 (2021).
34. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020).
35. Nichol, A. Q. & Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171 (PMLR, 2021).
36. Bakas, S. *et al*. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. *The Cancer Imaging Archive* (2017).
37. Bakas, S. *et al*. Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. *The Cancer Imaging Archive* (2017).
38. Bakas, S. *et al*. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data* **4**, 170117 (2017).
39. Bakas, S. *et al*. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv:1811.02629* (2018).
40. Menze, B. H. *et al*. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).
41. Baid, U. *et al*. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv:2107.02314* (2021).
42. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*: *18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
43. Liu, Z. *et al*. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
44. Hedlund, J., Eklund, A. & Lundström, C. Key insights in the AIDA community policy on sharing of clinical imaging data for research in sweden. *Scientific Data* **7**, 331 (2020).
45. Akbar, M. U. & Eklund, A. Synthetic brain tumor images from GANs and diffusion models. *AIDA datahub* https://doi.org/10.23698/aida/synthetic/brgandi (2023).
46. Borji, A. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding* **179**, 41–65 (2019).
47. Barratt, S. & Sharma, R. A note on the inception score. *arXiv:1801.01973* (2018).
48. Dikici, E., Bigelow, M., White, R. D., Erdal, B. S. & Prevedello, L. M. Constrained generative adversarial network ensembles for sharable synthetic medical images. *Journal of Medical Imaging* **8**, 024004–024004 (2021).
49. Mei, X. *et al*. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence* **4**, e210315 (2022).
50. Rajotte, J.-F. *et al*. Synthetic data as an enabler for machine learning applications in medicine. *Iscience* **25** (2022).
51. Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H. & Langlotz, C. P. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* **295**, 675–682 (2020).
52. Dar, S. U. H. *et al*. Investigating data memorization in 3D latent diffusion models for medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (pp. 56-65). Cham: Springer Nature Switzerland (October, 2023).
53. Isensee, F., Jäger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. Automated design of deep learning methods for biomedical image segmentation. *arXiv:1904.08128* (2019).
54. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
55. Cirillo, M. D., Abramian, D. & Eklund, A. What is the best data augmentation for 3d brain tumor segmentation? In *2021 IEEE International Conference on Image Processing (ICIP)*, 36–40 (IEEE, 2021).
56. Toolbox, O. S. S. & https://github.com/open-mmlab/mmsegmentation. 202, B. Mmsegmentation contributors. mmsegmentation: Openmmlab semantic segmentation.

Acknowledgements
Training several of the generative models and all the segmentation networks was performed using the supercomputing resource Berzelius (752 Nvidia A100 GPUs) provided by the National Supercomputer Centre at Linköping University, Sweden. It was donated by the Knut and Alice Wallenberg foundation. This research was supported by the ITEA/VINNOVA project ASSIST (Automation, Surgery Support and Intuitive 3D visualization to optimize workflow in IGT SysTems, 2021-01954), LiU Cancer, VINNOVA AIDA (2021-01420), and the Åke Wiberg foundation (M22-0088). Ida Blystad was supported by a research grant from the Wallenberg Center for Molecular Medicine as an associated clinical fellow.

Author contributions
M.U.A. trained all the generative models, the Swin transformer segmentation networks, generated synthetic images and calculated F.I.D. and I.S. metrics. M.L. trained all the U-Net segmentation networks and calculated the segmentation metrics. I.B. performed the qualitative evaluation. A.E. drafted the manuscript and contributed with conceptualization, supervision and funding. All authors contributed to the interpretation of the results, have revised and edited the manuscript and approved the submitted version.

Funding
Open access funding provided by Linköping University.

Competing interests
AE has previously received graphics hardware from Nvidia. The other authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.