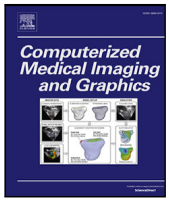
Multi-modal brain tumor segmentation via conditional synthesis with Fourier domain adaptation

Computerized Medical Imaging and Graphics 112 (2024) 102332

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



Multi-modal brain tumor segmentation via conditional synthesis with Fourier domain adaptation

Yasmina Al Khalil [a,*], Aymen Ayaz [a], Cristian Lorenz [b], Jürgen Weese [b], Josien Pluim [a], Marcel Breeuwer [a,c]

[a] Biomedical Engineering Department, Eindhoven University of Technology, Eindhoven, The Netherlands
[b] Philips Research Laboratories, Hamburg, Germany
[c] Philips Healthcare, Best, The Netherlands

ARTICLE INFO

Keywords:
Brain tumor segmentation
Image synthesis
Frequency domain adaptation

ABSTRACT

Accurate brain tumor segmentation is critical for diagnosis and treatment planning, whereby multi-modal magnetic resonance imaging (MRI) is typically used for analysis. However, obtaining all required sequences and expertly labeled data for training is challenging and can result in decreased quality of segmentation models developed through automated algorithms.

In this work, we examine the possibility of employing a conditional generative adversarial network (GAN) approach for synthesizing multi-modal images to train deep learning-based neural networks aimed at high-grade glioma (HGG) segmentation. The proposed GAN is conditioned on auxiliary brain tissue and tumor segmentation masks, allowing us to attain better accuracy and control of tissue appearance during synthesis. To reduce the domain shift between synthetic and real MR images, we additionally adapt the low-frequency Fourier space components of synthetic data, reflecting the style of the image, to those of real data. We demonstrate the impact of Fourier domain adaptation (FDA) on the training of 3D segmentation networks and attain significant improvements in both the segmentation performance and prediction confidence. Similar outcomes are seen when such data is used as a training augmentation alongside the available real images. In fact, experiments on the BraTS2020 dataset reveal that models trained solely with synthetic data exhibit an improvement of up to 4% in Dice score when using FDA, while training with both real and FDA-processed synthetic data through augmentation results in an improvement of up to 5% in Dice compared to using real data alone. This study highlights the importance of considering image frequency in generative approaches for medical image synthesis and offers a promising approach to address data scarcity in medical imaging segmentation.

1. Introduction

Accurate and consistent brain tumor segmentation is crucial for diagnosis, treatment planning and post-treatment assessment (Chen et al., 2017). Gliomas are a prevalent type of brain tumors, further divided into different tumor grades based on their underlying histology and molecular characteristics, whereby the most commonly studied ones include low-grade (LGG) and more aggressive high-grade (HGG) gliomas. Magnetic resonance imaging (MRI) is most commonly used for glioma diagnosis, providing the ability to extract complementary information from multiple sequences to distinguish and assess the key tumor components, such as the necrotic and non-enhancing region (NCR/NET), the peritumoral edema (ED) and the enhancing (ET) region. Typical sequences utilized include T1-weighted (T1w), T2-weighted (T2w), contrast-enhanced T1-weighted (T1ce) and Fluid Attenuation Inversion Recovery (FLAIR) images (Li et al., 2018; Zhou et al., 2019).

The current gold standard in brain tumor segmentation is manual tracing by professional radiologists (Işın et al., 2016; Zhao et al., 2019). However, manual segmentation is a tedious, labor-intensive and subjective process, leading to inter-expert variability and questionable accuracy. With the recent rise of machine and deep learning in medical image analysis, many automated segmentation algorithms have been proposed in the literature (Işın et al., 2016; Zhao et al., 2019). Nonetheless, automated brain tumor segmentation remains a challenging task, largely due to appearance and shape heterogeneity

* Corresponding author.
E-mail addresses: y.al.khalil@tue.nl (Y.A. Khalil), a.ayaz@tue.nl (A. Ayaz), cristian.lorenz@philips.com (C. Lorenz), juergen.weese@philips.com (J. Weese), j.pluim@tue.nl (J. Pluim), m.breeuwer@tue.nl (M. Breeuwer).

https://doi.org/10.1016/j.compmedimag.2024.102332
Received 4 July 2023; Received in revised form 31 October 2023; Accepted 13 December 2023
Available online 11 January 2024

of tumor lesions (Wang et al., 2019; Takahashi et al., 2021; AlBadawy et al., 2018).

While ample research in the field is focused on architectural changes to improve the segmentation performance of existing methods, these bring up only a small fraction of improvement, as all such algorithms mainly rely on the data available for training (Wadhwa et al., 2019; Magadza and Viriri, 2021). Research in other domains of computer vision and image processing has already established the need for attaining large datasets to train robust and generalizable deep learning models (LeCun et al., 2015).

However, acquiring cross-institutional heterogeneous data is a challenging task, made even more difficult by patient privacy and restrictive data sharing policies (Nalepa et al., 2019). To overcome these limitations, data augmentation has been introduced to artificially increase variation in existing data, where classical approaches of transforming the shape and appearance of images have now become a standard.

However, classical approaches rarely cover the extensive variations typically seen in MR images, often producing very similar or correlated samples, not beneficial to algorithm training (Shin et al., 2018).

Recent developments in generative adversarial networks (GANs) have enabled the generation of new images from both labeled and unlabeled original images, with various ways of injecting variations in the generation process (Chlap et al., 2021; Chen et al., 2022). Despite promising results achieved in the domain of medical image synthesis, considerable distortions are still reported when considering the frequency information of the generated data (Schwarz et al., 2021). In fact, most GAN-based synthesis approaches operate in image space only, which is not sufficient to adequately capture the low-level features influencing the contrast, texture and large-scale content of images during training. This leads to potentially sub-optimal results in a number of downstream tasks, such as segmentation (Zhang et al., 2022).

This study aims to achieve several objectives. First, we explore the feasibility of training segmentation models solely with fully synthetically generated data and evaluate the implications of relying on synthetic images. We perform the same analysis on utilizing synthetically generated images for augmentation, but we aim to determine the optimal level of variation and the required amount of data necessary to continually enhance the segmentation performance. By addressing these aspects, we aim to provide insights into the key factors influencing the effectiveness of augmentation with synthetic images for improving brain tumor segmentation. Moreover, we analyze the impact of introducing tumor shape variations during the synthesis process on the performance of the segmentation model and assess whether additional variations improve training outcomes. The study also delves into the effects of aligning the frequency components of synthetic images with those of real images during model training. Recent findings have indicated that low-level features, such as style, and semantic information in images can be captured through the low-frequency amplitude and phase components of the Fourier transform, respectively (Yang and Soatto, 2020; Yang et al., 2020). Therefore, we study whether conditional GANs effectively capture frequency-related image characteristics and assess the implications for model training and performance when using GAN-generated images.

2. Related work

Initial approaches focused on generating images using noise-to-image structures, originating from the vanilla GAN (Goodfellow et al., 2014) and its variants and synthesizing images from one-dimensional vectors (Frid-Adar et al., 2018). However, these methods are not able to perform a pixel-wise matching between two images, crucial for accurate representation of varying image contrasts in MR imaging. A popular approach is to take advantage of image-to-image translation frameworks, utilizing models such as pix2pix to generate the desired contrast (Isola et al., 2017). However, pix2pix and similar models require paired data, which is typically expensive to obtain for medical images. To overcome this limitation, the cycleGAN (Lei et al., 2019; Zhu et al., 2017; Hiasa et al., 2018) was developed for unpaired cross-modality synthesis and translation. However, unpaired image-to-image translation approaches still require the presence of two or more modalities during training. This can be alleviated by methods that focus on learning a mapping function between a semantic label and a specific modality (Mok and Chung, 2018), whereby conditional GANs have become increasingly popular.

Utilizing GAN-based synthetic images for the purpose of data augmentation was first demonstrated in Shin et al. (2018) for brain tumor segmentation using the pix2pix GAN model (Isola et al., 2017), conditioned on brain and tumor masks with additional tumor variations achieved by altering the input label maps. Qasim et al. (2020) used a spatially-adaptive de-normalization (SPADE) GAN model (Park et al., 2019) to synthesize tumor images for data augmentation, demonstrating an improved performance for tumor segmentation. The model is conditioned on both the local and global information to mitigate the global class imbalance problem. However, there is no additional augmentation to increase variability in the synthetic data.

While generative adversarial models are now widely utilized to synthesize images for a wide array of medical imaging modalities, we rarely see such methods incorporated in clinical practice. Although GANs enable us to incorporate invariance and robustness into the training of deep learning-based models in regards to not only affine transformations, but also variations in tissue shape and appearance, the convergence of adversarial training and the existence of its equilibrium point are still unresolved issues (Nalepa et al., 2019). Moreover, studies continuously show that generators tend to produce multiple very similar examples, which cannot serve as means to improve the generalization of downstream tasks, known as the mode collapse problem (Wang et al., 2017). A wide range of studies demonstrate that synthetic images are still not of adequate quality to train models that perform at the same level as their counterparts trained with real medical images. This implies the existence of an intrinsic domain shift between real and synthetic images. Recent studies report the evidence of considerable distortions in frequency information of synthetic data, especially at higher frequencies, despite the observed structural similarities with real data (Dzanic et al., 2020; Singh et al., 2021). Some of this discrepancy in Fourier components have been assigned to specific architectural components of neural networks, such as regularization strategies, up-sampling (Durall et al., 2020), as well as linear dependencies in the filter spectrum of convolutional layers (Khayatkhoei and Elgammal, 2022). A limited amount of studies (Tajmirriahi et al., 2022; Singh et al., 2021; Mizutani et al., 2016) have addressed these issues in the domain of medical imaging, where frequency components drive the acquisition and reconstruction of images.

3. Material and methods

3.1. Data

To train and test the proposed pipeline, we make use of the BraTS2020 dataset (Bakas et al., 2018), consisting of 369 clinically acquired, pre-operative multi-modal MRI scans of HGG (N = 293) and LGG (N = 76) tumors. A set of T1w, T2w, T2-FLAIR and T1ce contrasts are acquired per subject across multiple institutions (19), scanners and clinical protocols. Acquisition from variable sources resulted in images with differences in intensities, contrast, acquisition resolution and plane, artifacts and quality. All images are co-registered, interpolated to the same 3D resolution and skull-stripped. The publicly available training set is accompanied with manual, expert tumor annotations of ET, ED and NCR/NET tissue. Due to the unbalanced tumor grade distribution, we focus this work on HGG tumors only (N = 293). We split the available training data into two sets, for training the synthesis module and segmentation networks, respectively. In addition, we split

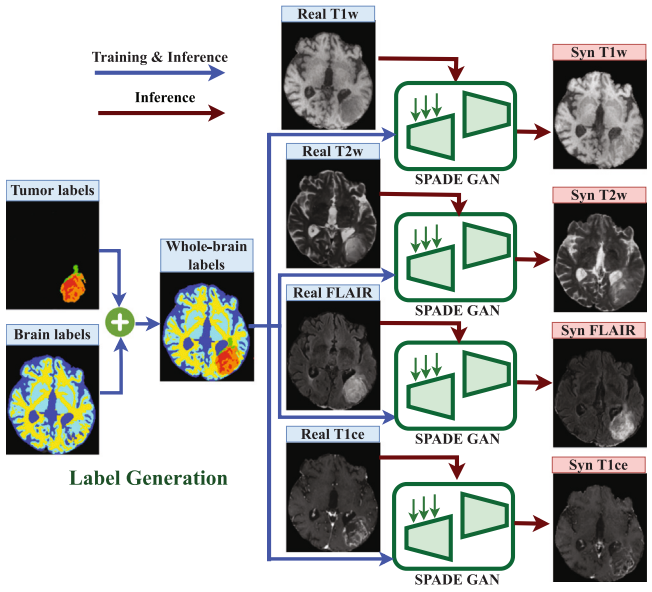


Fig. 1. An overview of the proposed methodology consisting of a comprehensive label generation part for conditional image synthesis using SPADE GANs for variable multi-modal brain tumor MR image generation. Six class labels generated covering full brain and HGG tumors along multi-modal MR tumor data are used for training along with their respective T1w, T2w, FLAIR and T1ce sequences.

the latter set of data into training and testing, as neither the testing nor the validation set is not publicly provided by the BraTS2020 challenge.

In total, we utilize 100 sequences per contrast to train each synthesis module (four modules to generate four MRI contrasts), while we allocate a separate set of 150 sequences for training all segmentation models in this work. Ground truth labels from the same 150 sequences are further deformed and used to generate additional synthetic sequences, for a total of 750 synthetic images per contrast. The remaining 43 sequence sets are allocated for testing of the trained segmentation models.

3.2. Conditional image synthesis

In this study, we focus on the utilization and assessment of synthetic images generated by conditional GANs. To be more precise, the proposed synthesis approach is conditioned on auxiliary brain tissue and tumor segmentation masks. Moreover, we stratify the training of synthesis models per tumor grade and focus on training a GAN model for HGG synthesis. Our previous experiments show that GANs trained on mixed-grade images tend to generate unrealistic contrasts, particularly when it comes to specific tumor regions (Khalil et al., 2022). This is additionally emphasized by the imbalanced training set, where examples of HGG tumors largely outnumber the LGG ones.

The overall synthesis pipeline consisting of label generation and conditional image synthesis is shown in Fig. 1. We utilize a supervised, mask-guided image generation technique that employs spatially adaptive denormalization (SPADE) layers (Park et al., 2019), reinforcing semantically-consistent image synthesis. The main advantage of this approach is provided by the utilization of SPADE layers, designed to inject information from the available segmentation maps throughout the network, guiding the generator to correctly learn the translation between tissue classes and their realistic appearance in real MR images. This is important for learning the textual features (style) of separate tissue classes and preventing information loss when passed through multiple convolutional layers. We opt for SPADE GANs as our previous work in Amirrajab et al. (2022) confirms the superiority of the SPADE generator in synthesizing high quality contrast compared to other popular models. Our experiments further show that utilizing comprehensive tissue labels, i.e. obtaining segmentation maps of all major structures in the image field of view (FOV) leads to improved quality and realism of the generated images. This allows us to produce realistic and consistent tissue contrast not only across tumorous regions, but also across all visible brain tissue, compared to similar approaches available in the literature (Qasim et al., 2020). Thus, in addition to expert annotations of tumor tissue (ET, ED and NCR/NET) available in the BraTS2020 dataset, we obtain a set of whole-brain labels of white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). These labels are generated on BraTS2020 T1w MR images for all HGG subjects, obtained using a model-based brain segmentation approach (Wenzel et al., 2018) combined with a multi-scale fully convolutional architecture for processing images at different scales and FOVs (Brosch and Saalbach, 2018). The generated labels are combined with the available tumor labels to obtain a six-class input label map per subject.

Separate 2D SPADE GANs are trained for HGG synthesis per modality, using multi-modal structural MR data along with the combined six-class label maps. All images undergo pre-processing through in-plane center cropping and removal of empty slices (in transversal direction) lacking brain tissue. Moreover, we utilize a matrix size of 192×192 with 128 axial slices and an isotropic resolution of 1 mm for both images and labels during training. All models are trained using the Adam optimizer with a learning rate of 0.0002 and a batch size of 12 on three NVIDIA TITAN Xp GPUs, without any data augmentation. All other training parameters, architecture and losses remain the same as in Park et al. (2019). In total, four separate 2D SPADE GANs are trained on axial slices of T1w, T2w, T1ce, and FLAIR brain tumor MR images, respectively. At inference time, all models require a set of combined tumor and brain tissue labels at the input. The same set of unseen labels derived from a real MRI is fed to all four image synthesis networks for obtaining T1w, T2w, T1ce, and FLAIR contrasts, representing the provided tissue mask.

To increase synthetic image variability, we modify input label maps by changing the tumor's shape and size using three deformations: elastic transformations, dilation, and erosion. Random elastic deformations are applied using seven control points along each dimension of the coarse grid, with a maximum displacement of 10 voxels for plausible deformations. These deformations are applied to the complete set of labels, including the whole brain and tumor. Morphological deformations, such as dilation and erosion, are only applied to tumor class labels using spherical structuring elements of different diameters. This results in the enlargement and shrinking of different tumor regions. The deformed tumor class labels are then combined with the original brain labels to create multiple plausible combinations, including the enlarged NCR/NET, ET, and ED, as well as the shrunk ED with enlarged ET. Additionally, elastic deformations are applied to NCR/NET, ED, ET, GM, WM, and CSF.

Our primary goal is to generate realistic MR images with diverse tumor shapes and sizes to enhance segmentation accuracy by capturing textural and contrast changes in tumor regions. However, modeling the intricate anatomy and appearance of various tumor tissues is challenging and not the main objective of our work. We recognize that elastic and morphological deformations can result in highly unrealistic shapes, so we ensure that the generated shapes align with volumetric variations documented in the literature (Dang et al., 2022), thereby maintaining plausibility. Additionally, we take into account the relationship between neighboring tissues, considering the structural changes observed in different patients. For instance, we acknowledge that edema, which is typically diffuse and less well-defined compared to the enhancing tumor, often extends beyond the boundaries of the tumor itself, encompassing a larger area of the brain tissue. Furthermore, the NCR/NET tumor can exhibit a wide range of sizes, spanning from small focal areas to larger extensive regions. To capture this variability, we generate both prominent non-enhancing regions that encompass a significant

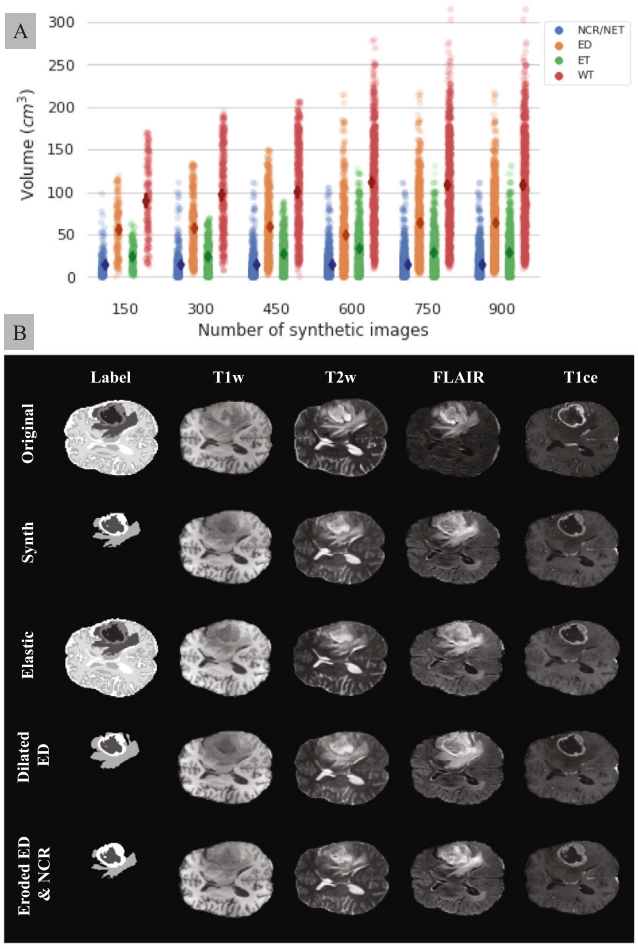


Fig. 2. (A) Range of volumes per tumor region and the whole tumor represented by different subsets of generated synthetic data, where variation in tumor shape and size is encouraged through transforming brain and tumor labels at the input of the synthesis module. Examples of tumor and whole brain transformations for a single subject across different contrasts, compared to the original tumor/brain shape, is shown in (B).

proportion of the tumor volume and small focal areas, ensuring that our generated images reflect the full spectrum of tumor sizes.

Examples of different transformations applied to labels and the generated synthetic images can be observed in Fig. 2. Fig. 2 also shows the change in both brain tissue and shape volumes as various subsets of generated synthetic data are selected for training. We observe a trend of increasing tumor volume variation with larger datasets, but no significant increase in shape and volume diversity is found for subsets with over 600 images. Therefore, we experiment with different subsets of training data to evaluate whether they contain sufficient variation to support the training process.

3.3. Fourier domain adaptation

While we produce synthetic images of realistic contrast per modality and tumor grade, there are several observed issues with the appearance of the generated images compared to real MRIs. First, while GANs can produce diverse samples capturing the statistical distribution inherent to the training data, the generated images often exhibit reduced contrast and a narrower intensity range, resulting in a perceptually lower dynamic range compared to real images. Secondly, GAN-based synthesis methods, currently constrained to image space, face challenges in accurately reproducing textures and fine details. This limitation largely arises from a spectral bias favoring low-frequency signals over high-frequency components, resulting in the loss of fine-grained information and texture details in the generated images (Li et al., 2023).

Operating in k-space or on frequency components of images has only recently been recognized as an efficient, but simple domain shift adaptation strategy when tackling images coming from different sources or domains (Kong and Shadden, 2020; Liu et al., 2021). Accordingly, we adapt the Fourier domain adaptation (FDA) technique in Yang and Soatto (2020) to 3D images and swap the low-frequency amplitudes of synthetic images with those of real images belonging to the same modality (see Fig. 3). As amplitudes of the low-frequency spectrum components mostly contain the information about low-level image characteristics, typically capturing the strength of global or large-scale variations in overall brightness, intensity distribution, contrast and texture, this approach helps with aligning the domain shift between the distributions of synthetic and real images. GANs may struggle to precisely capture the complexity and variability of these low-level characteristics present in real images, leading to discrepancies between the synthetic and real domains.

If real MR image counterparts are available for synthetic images during synthesis, they can be directly used for the FDA. However, when there is a limited number of image contrast examples compared to a large set of brain tissue and tumor labels, caution should be exercised when applying FDA. Randomly swapping spectral components between images can lead to visual artifacts, particularly if the images differ significantly in semantics. To ensure semantic similarity before applying FDA, metrics like the spectral residual similarity index (SR-SIM) (Zhang and Li, 2012) can be employed. In our approach, we carefully select image pairs based on their shared brain tissue labels, prior to applying any morphological or elastic deformations. In other words, the selection of the original real image to which the synthetic image will be adapted to using FDA is based on the fact that its brain tissue and tumor labels were morphologically or elastically deformed to generate that specific synthetic image. Here, we consider the corresponding synthetic images the most similar semantically to real images from which the original labels, prior to the deformation step, are extracted.

We perform the alignment in the Fourier domain as follows. For simplicity, let $\mathcal{F}_a, \mathcal{F}_p : \mathbb{R}^{HxWx1} \rightarrow \mathbb{R}^{HxWx1}$ be the amplitude and phase components of the Fourier transform $\mathcal{F}$ of a 2D gray-scale, single channel image I:

$$\mathcal{F}(I)(m,n) = \sum_{h,w} I(h,w) e^{-j2\pi(\frac{h}{H}m + \frac{w}{W}n)}, j^2 = -1, \tag{1}$$

implemented using the FFT algorithm (Frigo and Johnson, 1998). Using the inverse Fourier transform, $\mathcal{F}^{-1}$, image phase and amplitude can be mapped back to image space. Furthermore, a low-frequency amplitude cut-out window, M_β, whose value is zero for the region where $\beta \in (0,1)$, is defined as:

$$M_\beta(h,w) = \mathbb{1}_{(h,w)} \in [-\beta H : \beta H, -\beta W : \beta W], \tag{2}$$

where β represent the size of the swapping window and does not depend on image size and resolution. Thus, given two paired images $I^s \in D^s$ and $I^t \in D^t$ from source and target domains, respectively, FDA is defined as:

$$I^{s \rightarrow t} = \mathcal{F}^{-1}([M_\beta \circ \mathcal{F}_a(I^t) + (1 - M_\beta) \circ \mathcal{F}_a(I^s), \mathcal{F}_p(I^s)]), \tag{3}$$

where the low frequency part of the source image amplitude $\mathcal{F}_a(I^s)$ is replaced by the one extracted from the target image, while its phase component $\mathcal{F}_p(I^s)$ remains unaltered. Such modified spectral representation is mapped back to image $I^{s \rightarrow t}$. The procedure is visualized in Fig. 3.

For $\beta = 0$, the adapted image $I^{s \rightarrow t}$ will remain the same as the original source image I^s. On the other hand, with $\beta = 1$, the amplitude of I^s will be completely replaced by that of I^t. Moreover, as the value of β increases to 1, the amount of visible artifacts tends to increase. Thus, for each of the four sequences used in this study, we experimentally

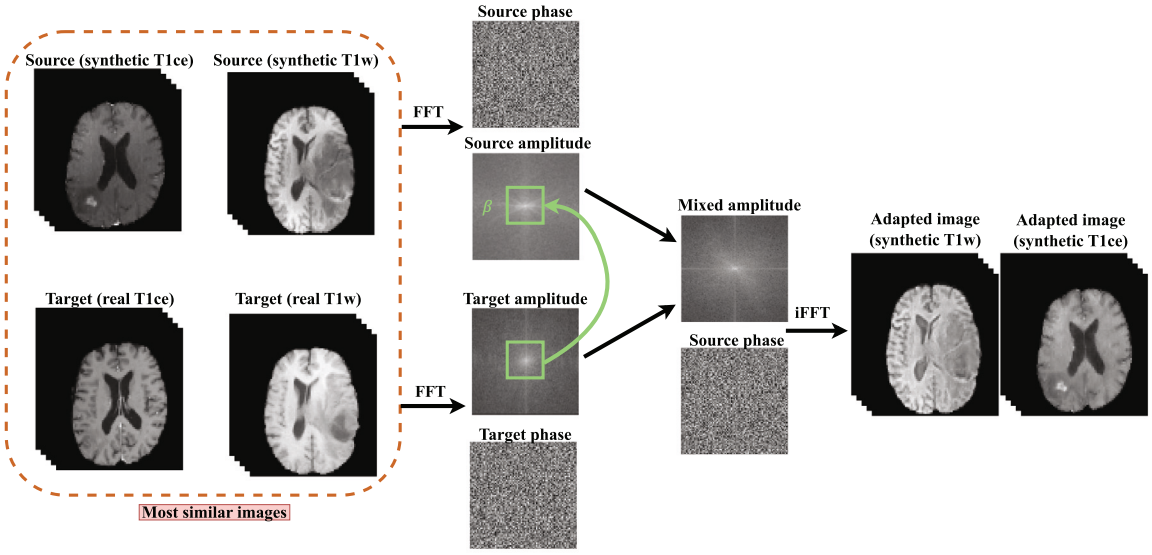


Fig. 3. Fourier domain adaptation (FDA) applied to synthetic images to address the distortions in the frequency spectrum.

determine the value of β that produces images similar in appearance to the real target MR images, while avoiding the generation of extreme artifacts (see Appendix C, Table 1). In particular, the values of β for the adaptation of T1w, T2w, FLAIR and T1ce images are chosen to be 0.1, 0.001, 0.03 and 0.001, respectively. We generally observe that higher values of β distort the T2w and T1ce sequences significantly in some cases, while they work better for T1w and FLAIR images.

Note that before FDA, all images are normalized to intensities from [0,1], as well as resampled to an isotropic resolution of 1 mm. After FD adaptation, we apply contrast stretching, rescaling the image intensity levels to include all intensities that fall within the 2nd and 98th percentile.

3.4. Segmentation

3.4.1. Network architecture and training

To perform the experiments in this study, we adapt a 3D nnU-Net (Isensee et al., 2021) model for a multi-class and multi-channel segmentation task with several modifications to improve the generalization of the model to a variety of sequences utilized in this study. We replace the standard instance normalization layers of the baseline nnU-Net with batch normalization, use Leaky ReLUs and introduce heavier data augmentation compared to the standard pipeline. These include image scaling ($p = 0.3$) with a scaling factor in the range of [0.7–1.4], random rotations within ±60 degrees ($p = 0.7$), random horizontal and vertical flips ($p = 0.3$) and elastic transformations ($p = 0.3$). Moreover, we apply intensity transformations in the form of gamma correction ($p = 0.3$) with the gamma factor ranging within [0.5–1.6], additive brightness transformations ($p = 0.3$) with the brightness factor varying within [0.7–1.3] and multiplicative brightness ($p = 0.3$) with a mean of 0 and standard deviation of 0.3.

Each imaging sequence is fed separately per channel, forming an input of size 4 × H × W × D, where 4 denotes the number of channels or sequences used, while H, W, D stand for image height, weight and depth, respectively. The input patch size is selected to be 96 × 160 × 128 with a batch size of 2. We modify the data-loader to ensure that each batch consists of one synthetic and one real image, to avoid over-fitting on synthetic images. A total of five down-sampling operations are performed, with an initial number of convolutional kernels set to 32. We use a combination of Dice and cross-entropy loss, optimized using Adam for stochastic gradient descent, with an initial learning rate of 10^{-3} and a weight decay of $3e^{-5}$. The loss operates on the three target labels — edema, necrosis and enhancing tumor. During training, the learning rate is reduced by a factor of 5 if the validation loss has not improved by at least 5×10^{-3} for 50 epochs. We train all models for a maximum of 1000 epochs, where early stopping is applied when the learning rate drops below 10^{-6}. Moreover, all models are optimized and trained under a 5-fold cross-validation set-up using all training data.

3.4.2. Post-processing

As a post-processing step, we remove the false positive predictions of the enhancing tumor in cases when the predictions are below a certain threshold and replace them by the necrotic tissue. The optimal threshold is selected using the training set cross-validation approach, according to the best mean Dice score of the enhancing tumor region. This is followed by a connected component analysis on the predicted labels, where we remove all but the largest connected component per class.

3.5. Evaluation

To quantitatively assess the quality of synthesized images and the effect of FDA on synthetic data, we utilize reference-based image quality metrics, which include the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR) and the mean squared error (MSE). We additionally include a measurement of SR-SIM, which uses spectral residual visual saliency maps as a feature to compute the local similarity maps between two images (Zhang and Li, 2012). SR-SIM is designed based on the assumption that an image's visual saliency map has a close relationship with its' perceptual quality. The quantitative assessment of image quality is done on a small subset of 50 synthetic images generated using the original labels from real multi-modal images from the segmentation training set (see Section 3.1), which serve as reference images during the computation of image quality metrics. The selected synthetic images are adapted using the FDA approach to one of the remaining 100 original real images from the assigned segmentation training set, whereby we use SR-SIM to match the most structurally and semantically similar images.

To assess the performance of the proposed approach, we perform a quantitative evaluation in terms of the standard segmentation metrics reported in the literature — Dice score, Hausdorff distance (HD), sensitivity and specificity. To understand the impact of augmentation and training with synthetic data on model confidence and prediction quality, we utilize Monte Carlo (MC) dropouts for uncertainty mapping (Gal and Ghahramani, 2016). We add "infer-dropouts" to each trained

model during inference and insert them after each contraction and expansion block with dropout rates of {0.01, 0.025, 0.05, 0.1, 0.2, 0.5}. Along with the standard inference without any dropout, this results in six tests per model, using the specified dropout rates on the same test data, without changing any model weights or parameters. From the predicted softmax probabilities, we compute a final uncertainty map as a pixel-wise variance across all predictions.

4. Experiments

This study aims to evaluate the efficacy of multi-modal synthetic brain MR images in localizing and segmenting high-grade gliomas in real MR sequences, while exploring the impact of the number of synthetic images and diversified tumor shapes during training on segmentation performance. The second major aspect of the proposed study is addressing the perceived gap between the frequency spectrum of real images and those produced by GANs, as discussed in 3.3.

We utilize 150 multi-modal sequences per patient, synthesized from the BraTs training set, and an additional 150 sequences per deformation strategy (see Section 3.2), for a total of 750 images with varying tumor shapes. Our investigation focuses on the following aspects:

1. We evaluate the **segmentation performance of models trained on synthetic sequences**, whereby multiple models for whole-tumor detection and segmentation are trained on a gradually increasing number of synthetic sequences generated from deformed tumor shapes. These models are then compared to a baseline model trained on 150 real multi-modal MR sequences.
2. We explore the **effects of Fourier domain adaptation (FDA) on the performance of models trained with synthetic images**. FDA is applied to all generated images, followed by re-training the aforementioned models using the adapted images.
3. We additionally evaluate the **quality of synthetic images with and without the application of FDA** using several different computational metrics that reflect their perceptual and structural quality compared to their multi-modal real image counterparts.
4. We assess the impact of **utilizing synthetic multi-modal images for augmentation**. This involves gradually augmenting 150 real multi-modal sequences with an additional set of 750 sequences, which exhibit diverse tumor shapes. We investigate the role of frequency components in addressing the domain shift between real and synthetic data.

5. Results

Training with synthetic data: Fig. 4(a) and (b) show the segmentation performance of multi-modal segmentation models trained on synthetic sequences generated in this work in terms of Dice and 95th percentile Hausdorff Distance (HD95), respectively, whereby the amount of data is gradually increased during training. As the amount of data increases, we note an improvement in model performance, which saturates at around 600 synthetic images used for training. Further expansion of the training set does not lead to additional gains in performance, as observed by the Dice and HD95 scores. We hypothesize that enlarging the training set further does not offer any additional variation in shape or contrast for the network to uncover and learn new information. The findings in Fig. 2 A indicate that training sets comprising more than 600 synthetic images exhibit restricted diversity in the shapes and sizes of various tumor regions. Consequently, this limited diversity may not yield a substantial increase in valuable information for networks to capture, resulting in negligible performance improvements. This limitation could be partly influenced by the availability of real data with annotations for training the synthesis model and manipulating the anatomy through spatial transformations. Additionally, the nature of the transformations used, such as morphological and elastic transformations, constrains the number of plausible anatomical shapes that can be generated. Finally, utilizing a large number of synthetic images may lead to overfitting due to the presence of highly similar examples, negatively impacting the training process.

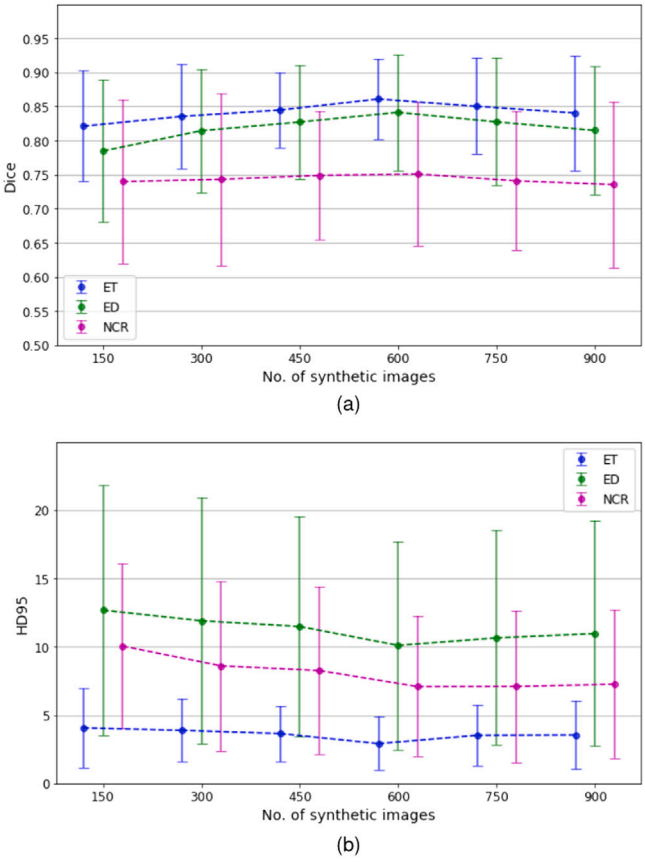


Fig. 4. Change in segmentation performance when training with a progressively increasing number of synthetic images in terms of the (a) Dice and (b) 95th percentile Hausdorff Distance. All scores are calculated across three different tumor regions: necrotic and non-enhancing tumor (NCR/NET), peritumoral edema (ED) and enhancing tumor (ET).

Effects of FDA: Considering the best performance is obtained using 600 synthetic images for training, we select this model for further analysis. Table 1 outlines the performance of the segmentation model trained solely on synthetic data in comparison to the baseline model, trained on real images. While across a number of tissues and performance metrics, the baseline model produces significantly better results (p-value < 0.01), the performance of the model trained on synthetic data can be comparable to training with real data, as well as to other results reported in the literature (Tiwari et al., 2020; Soomro et al., 2022). However, applying FDA to all synthetic images and retraining the model leads to significant performance improvements compared to training on unadapted synthetic data. The average Dice and precision scores are improved by 2%–3% compared to training without FDA, while we note an improvement of up to 12% in recall. In addition, the results of the model trained with FDA post-processed synthetic images do not significantly differ from those of the baseline model (p-value > 0.01) across most metrics. These results suggest that the frequency spectrum's domain shift is a contributing factor to why synthetic images generated through standard GAN-based methods cannot fully replace real data during training. This is in addition to the already extensively studied domain shift in the image domain.

While the above experiment suggests that FDA has a positive impact on a downstream segmentation task, it does not indicate whether FDA also improves the perceptual and structural quality of synthetic images.

Table 1
Segmentation performance of multi-modal segmentation whole-tumor networks trained on a varying number of synthetic sequences, with and without the application of FDA. All models are compared to the baseline, M_B, trained on original 150 training sequences in the BraTS dataset. We report the average Dice, 95th percentile Hausdorff distance (HD95), sensitivity and precision scores, along with standard deviation. Bold values indicate the best performing values per metric.

Methods	Dice				HD95			
	WT	NCR/NET	ED	ET	WT	NCR/NET	ED	ET
Baseline (N = 150)	0.921 (0.02)	**0.793 (0.21)***	**0.872 (0.12)**	**0.882 (0.11)**	**9.37 (6.31)***	5.53 (5.08)	**5.31** (4.89)*	2.87 (2.67)
M_Syn (N = 600)	0.897 (0.06)	0.745 (0.22)	0.831 (0.13)	0.846 (0.11)	12.03 (9.32)	7.29 (6.51)	10.13 (8.95)	3.16 (2.91)
M_Syn + FDA	**0.933 (0.04)***	0.774 (0.18)	0.863 (0.11)	0.872 (0.09)	9.87 (8.21)	**5.36 (5.31)**	7.81 (6.93)	**2.55 (2.43)**
	Precision				Recall			
Baseline (N = 150)	**0.971 (0.02)**	0.745 (0.25)	0.848 (0.16)	**0.972 (0.02)***	**0.934 (0.03)**	**0.842 (0.15)**	**0.911 (0.07)***	0.819 (0.17)
M_Syn (N = 600)	0.932 (0.05)	0.789 (0.26)	0.837 (0.17)	0.918 (0.06)	0.887 (0.07)	0.742 (0.27)	0.828 (0.12)	0.809 (0.18)
M_Syn + FDA	0.963 (0.03)	**0.833 (0.21)***	**0.862 (0.16)***	0.954 (0.04)	0.931 (0.04)	0.834 (0.12)	0.861 (0.11)	**0.872 (0.11)**

* Indicates any statistically significant result with respect to other models, according to the Wilcoxon signed-rank test for $p < 0.01$.

Thus, we randomly select a subset of 50 multi-modal images is from a training set of 150 images allocated for segmentation network training, comparing them with their synthetic counterparts with and without the application of FDA, derived from original brain tissues and tumor labels, ensuring a straightforward structural similarity assessment. In this experiment, the Fourier domain adaptation (FDA) has been performed by aligning the selected 50 images to one of the remaining 100 images from the training set using the SR-SIM measure.

We choose a realistic approach in the experiment by using available real images for FDA adaptation, instead of directly applying it to original real MRI counterparts of synthetic images. To mitigate the detrimental effects of spectral component swapping, source and target images are chosen based on semantic similarity using SR-SIM. Although SR-SIM is primarily for 2D images, we adapt it for a 3D context by calculating average similarity scores between scan slices. We repeat this selection and adaptation process across various MR modalities, allowing images from the same patient to match with different targets based on modality. We then compare the adapted to real MR images using standard measures like MSE, SSIM, and PSNR, assessing the FDA's efficacy in enhancing image quality. In addition, we report the SR-SIM as an indicator of spectral component similarity. This comprehensive process, including comparisons for non−adapted synthetic images, offers a valuable insights into FDA′s influence on perceptual and structural image quality. The summarized results are available in Table 2, focusing primarily on brain-related structures. The results indicate significant improvements across all metrics for T1w synthetic images after the application of FDA, while T2w and T1ce exhibit significant improvements according to SSIM and PSNR scores. Notably, the derived SR-SIM values of synthetic images reveal substantial distortions in their spectral components. Applying FDA is shown to significantly reduce these distortions across all modalities, whereby a higher SR-SIM value signifies enhanced structural and spectral similarity with reference to the original real image.

Augmentation with synthetic images: In Table 3, we can observe the results of augmenting the allocated training set of real multi-modal MR sequences with synthetic images, with and without the application of FDA post-processing before training. We report the best results across models trained with varying amounts of synthetic data (see Appendix A, Fig. A.1), which are obtained for a total of N = 600 synthetic sequences added to the available 150 real sequences. While we already achieve considerable improvement in the segmentation performance by augmentation with synthetic images (model **M_Aug**), utilizing FDA (**M_Aug + FDA**) leads to further improvements. In fact, across some tissues, such as NCR/NET and ED, we note significant (p-value < 0.01) improvements in Dice and precision scores, while similar results are also observed for the segmentation of the whole tumor across HD95 and recall values. Visual observation confirms that the application of FDA helps with the segmentation of smaller structures, such as the tumor core, necrotic region or the boundary areas of the peritumoral edema. Examples of segmentation maps derived from differently trained models are available in Fig. B.1 (Appendix B).

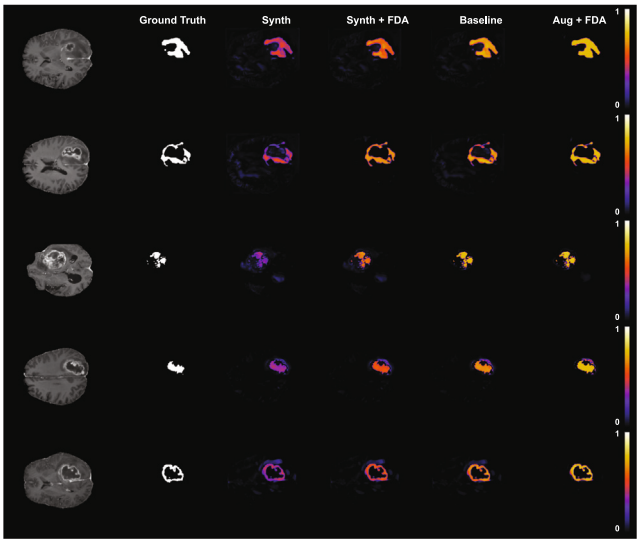


Fig. 5. Confidence or uncertainty maps of model predictions across different tumor regions plotted as normalized pixel-wise variance of infer-dropout predictions per model. We compare the baseline model, trained on all available real MR sequences to models trained entirely using synthetic data with and without FDA, as well as to a model augmented with FDA post-processed synthetic data. All variance values range between 0 and 1, where higher values (yellow) indicate an increase in prediction confidence of the model. Lower values (close to 0) indicate regions where the model is uncertain.

To further understand the impact of synthetic images on training, we plot uncertainty maps of pixel-wise prediction variance obtained from infer-dropouts described in Section 3.5. Fig. 5 displays uncertainty maps for various tumor regions, indicating the models' confidence levels when trained solely on synthetic data with and without FDA, as well as when using synthetic data in augmentation. We present only a small subset of examples, mainly those in which models trained on synthetic data display high uncertainty in predictions. The examples in Fig. 5 demonstrate a substantial decrease in model confidence when trained on synthetic data without FDA post-processing. However, applying FDA significantly improves model confidence, reducing under- or over-segmentation in these areas. Using FDA post-processed synthetic sequences for augmentation results in significant improvement over the baseline model trained solely on real sequences. Notably, there are significant gains in confidence at the tumor region borders and smaller structures, which are often challenging to delineate visually. Fig. 6 illustrates the probability distribution of certainty scores per tumor region for models trained on synthetic data with and without FDA, as well as for augmentation, evaluated across the entire testing set. Models trained on synthetic data without FDA exhibit the lowest certainty in predicting tumor regions, as supported by our previous observations and quantitative results. However, incorporating

Table 2
Mean, standard deviation, max and min of image quality metrics of HGG synthesized multi-modal MR images with and without FDA compared to real MRI. Mean values of any statistically significant improvements in the derived scores of Fourier domain adapted (FDA) synthetic images with respect to synthetic images without FDA are indicated in **bold**, according to the Wilcoxon signed-rank test for $p < 0.01$.

		T1w	T1w + FDA	T2w	T2w + FDA	FLAIR	FLAIR + FDA	T1ce	T1ce + FDA
SSIM	Mean	0.747	**0.823**	0.841	**0.857**	0.797	**0.832**	0.842	**0.855**
	Std	0.075	0.039	0.023	0.017	0.027	0.056	0.037	0.028
	Max	0.905	0.919	0.879	0.898	0.846	0.921	0.881	0..893
	Min	0.598	0.647	0.749	0.782	0.641	0.671	0.613	0.697
PSNR	Mean	15.643	**18.908**	20.122	**23.192**	20.906	22.121	22.893	**25.818**
	Std	3.025	2.004	1.717	1.271	1.694	1.682	1.763	1.801
	Max	21.456	27.542	23.937	26.322	24.359	27.571	25.861	27.921
	Min	9.159	12.811	16.926	18.541	15.138	16.381	15.482	18.754
MSE	Mean	0.033	**0.014**	0.007	0.006	0.009	**0.005**	0.006	0.004
	Std	0.016	0.012	0.003	0.003	0.004	0.003	0.003	0.002
	Max	0.121	0.059	0.021	0.011	0.031	0.018	0.031	0.014
	Min	0.007	0.002	0.004	0.002	0.004	0.002	0.003	0.002
SR-SIM	Mean	0.333	**0.679**	0.492	**0.741**	0.473	**0.723**	0.512	**0.759**
	Std	0.068	0.128	0.239	0.167	0.231	0.116	0.321	0.183
	Max	0.483	0.879	0.756	0.913	0.692	0.893	0.873	0.912
	Min	0.169	0.356	0.203	0.493	0.218	0.431	0.347	0.592

Table 3
Effect of augmentation using synthetic MR sequences (N = 600) with and without the application of FDA, compared to the baseline model trained on 150 real MR sequences of brain tumor images. The average DSC, HD95, precision and recall scores, along with their respective standard deviation values, are reported across the whole tumor (WT) and different tumor regions (NCR/NET, ED and ET). Bold values indicate statistically significant improvement w.r.t to the baseline.

Methods	Dice				HD95			
	WT	NCR/NET	ED	ET	WT	NCR/NET	ED	ET
Baseline (N = 150)	0.921 (0.02)	0.791 (0.21)	0.874 (0.12)	0.881 (0.11)	9.37 (6.31)	5.53 (5.08)	5.31 (4.89)	2.87 (2.67)
M_Aug (N = 750)	**0.945 (0.03)**	**0.808 (0.17)**	**0.897 (0.07)**	**0.892 (0.09)**	**8.11 (7.39)**	**4.26 (4.08)**	**2.93 (2.72)**	**1.85 (1.73)**
M_Aug + FDA	**0.952 (0.03)**	**0.821 (0.15)***	**0.911 (0.06)***	**0.903 (0.07)**	**7.85 (6.83)***	**3.39 (3.11)***	**2.52 (2.11)**	**1.61 (1.52)**
	Precision				Recall			
Baseline (N = 150)	0.965 (0.02)	0.753 (0.25)	0.853 (0.16)	0.971 (0.02)	0.925 (0.03)	0.841 (0.15)	0.912 (0.08)	0.819 (0.17)
M_Aug (N = 750)	0.971 (0.02)	**0.846 (0.18)**	**0.883 (0.12)**	0.978 (0.02)	**0.947 (0.03)**	**0.865 (0.12)**	**0.922 (0.06)**	**0.898 (0.08)**
M_Aug + FDA	**0.978 (0.01)**	**0.865 (0.11)***	**0.906 (0.08)***	**0.983 (0.01)**	**0.953 (0.02)***	**0.871 (0.11)**	**0.934 (0.04)***	**0.907 (0.06)***

* Marks significant improvements achieved by applying FDA compared to **M_Aug**, according to the Wilcoxon signed-rank test for $p < 0.01$.

FDA improves model performance, and using synthetic images for augmentation further increases model confidence compared to the baseline trained on real data only.

Effect on other segmentation architectures: In order to evaluate the consistency of the effects that the utilization of synthetic images with and without FDA has on the performance of segmentation networks in general, we compare the performance of the 3D nnU-Net models used in this study to other 3D state-of-the-art (SOTA) architectures proposed for multi-modal brain tumor segmentation. We perform the comparison under the same setup as reported in Tables 1 and 3, whereby a total of 600 synthetic images with and without FDA are utilized for training each network solely(**Syn** and **Syn + FDA**, respectively), as well as for training data augmentation (**Aug** and **Aug + Syn**, respectively). The total number of images used in the augmentation scenario is always 750, where the remaining 150 multi-modal images belong to real MRI segmentation training data. All baseline networks are trained only with this set of 150 multi-modal real images.

In order to provide a fair comparison, we try to train all networks in a consistent manner, minimizing the differences as much as possible between the training parameters and procedure. All networks are trained in a multi-modal manner, consisting of 4 $96 \times 160 \times 128$ input channels for each modality. If both synthetic and real images are used during training (augmentation experiments), the data-loader is modified to ensure that each batch consists of an equal number of synthetic and real images, to avoid over-fitting. Unless otherwise specified, all networks utilize a combination of Dice and cross-entropy loss, optimized using Adam with variable learning rates and weigh decay, as well as

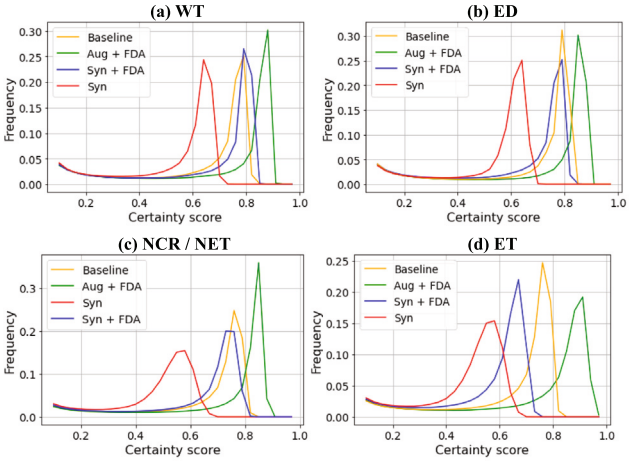


Fig. 6. Probability distribution plots of certainty scores per tumor region for models trained using synthetic data only with and without FDA, as well as the model augmented with FDA-processed images, compared to the baseline trained with real data. Regions evaluated include (a) the whole tumor (WT), (b) peritumoral edema (ED), (c) necrotic and non-enhancing tumor region (NCR/NET) and (d) the enhancing tumor (ET).

learning rate reduction based on the validation loss in a 5-fold cross-validation set-up. We further utilize a consistent data augmentation

set-up, as described in Section 3.4.1. Finally, early stopping is used for all models when the learning rate drops below 10^{-6}. The following SOTA architectures are evaluated:

- **3D U-Net** (Çiçek et al., 2016): a simple implementation of a 3D U-Net, trained end-to-end from scratch with a total of 4 encoder–decoder layers and a batch size of 2, the initial learning rate of 0.001 and a weight decay of $1e^{-5}$;
- **V-Net** (Milletari et al., 2016): an end-to-end fully convolutional 3D neural network with residual layers and the replacement of max-pooling operations with convolutional ones, trained using a batch size of 2, an initial learning rate of 0.0001 and a weight decay of 10^{-5};
- **Autoencoder regularization (AE Reg)** (Myronenko, 2019): a 3D encoder–decoder architecture with a variational autoencoder (VAE) added to the network to reconstruct the input images jointly with segmentation in order to regularize the shared encoder, trained using a combination of the Dice loss, L2 loss on the VAE branch and a VAE penalty term based on the KL divergence with an Adam optimizer and an initial learning rate of 0.0001, progressively decreased (as described in Myronenko, 2019) for a total of 300 epochs and a batch size of 1.
- **Hi-Net** (Qamar et al., 2021): a hyper-dense inception 3D U-Net trained using a combination of cross-entropy and Dice loss with an Adam optimizer, an initial learning rate of 3×10^{-5}, a batch size of 2 and a weight decay of $1e^{-5}$;
- **TransBTS** (Wang et al., 2021): a transformer-based encoder–decoder network trained with a batch size of 16 on 8 NVIDIA Titan Xp GPUs (12 GB) for 6000 epochs using a combination of Dice and cross-entropy loss with an Adam optimizer, an initial learning rate of 0.0004 and a weight decay rate of 10^{-5}, without any test time augmentation applied as in the original work.

Fig. 7 outlines the mean Dice scores and standard deviation per brain region across all networks, trained under five different set-ups. We observe a similar pattern as reported in Tables 1 and 3, where training with synthetic images incurs a drop in performance across all networks. However, applying FDA partially alleviates this decrease in performance, as evidenced also in the augmentation experiments. While both Hi-Net and TransBTS outperform other methods, the results obtained by utilizing the 3D nnU-Net optimized in this work consistently yield the highest results. Nevertheless, it is important to note that most methods in this experiment were utilized as-is, with limited time invested in optimization.

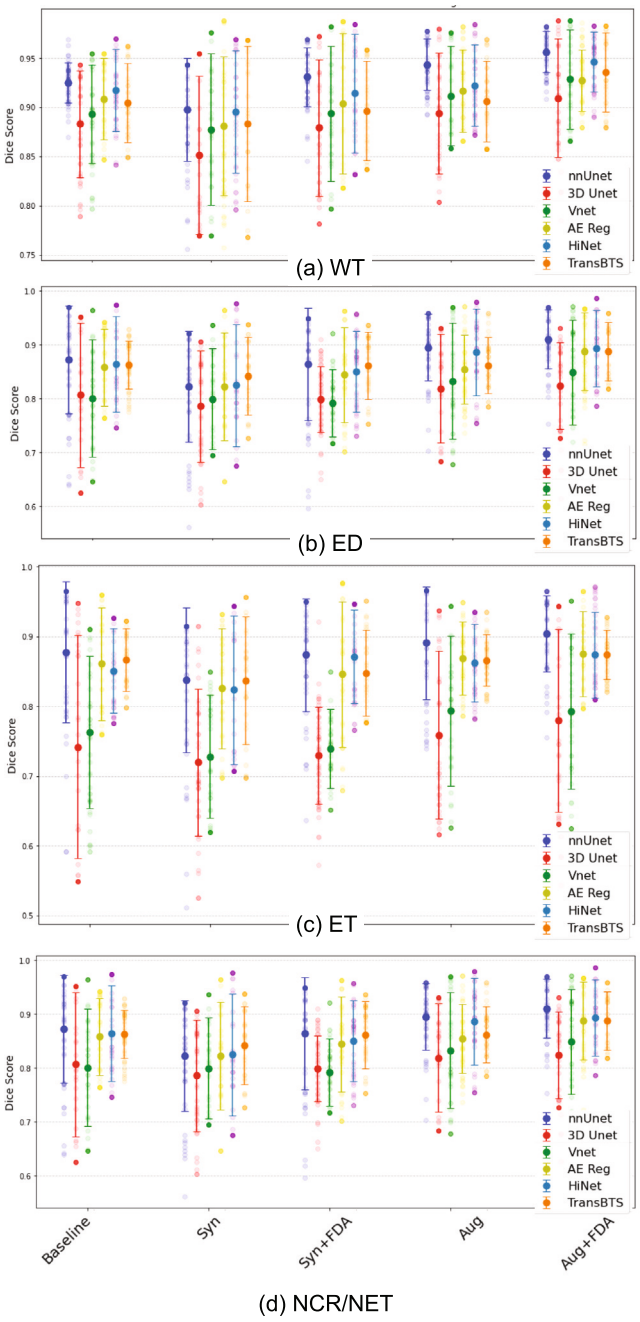


Fig. 7. Effect of utilizing synthetic multi-modal images for brain tumor segmentation with and without FDA for training and augmentation of different 3D segmentation architectures. All results are reported in terms of mean Dice scores and standard deviation across four different brain regions.

6. Discussion

In this work, we demonstrate the viability of utilizing fully synthetically generated brain MR sequences for brain tumor segmentation. We achieve this through a conditional architecture based on SPADE GANs, guided by fully semantic labels of the brain and tumor regions, trained to generate corresponding T1w, T2w, FLAIR and T1ce images per input label. Since the content pathway of the generative model is completely separated from the style pathway, we are able to condition the GANs on any labels, which we utilize to generate a large number of data with variations in tumor shape and size using elastic and morphological transformations. We further demonstrate that such highly variable GAN-based synthetic data can be used to train segmentation models with reasonable performance on the target distribution composed of real data. However, we still note a discrepancy in performance between models trained with synthetic data compared to training with real MR sequences, and show that it can be attributed to the disparity between the frequency spectra of images. More importantly, this disparity negatively impacts models augmented with synthetic data, limiting their performance on unseen real data. By utilizing Fourier domain adaptation and adjusting the low-frequency amplitudes of synthetic images to those of real images, we are able to partly account for this discrepancy and improve the performance of models trained on synthetic data. This is further confirmed through the experiments on uncertainty analysis using the MC dropout technique, demonstrating that using FDA post-processing on synthetic images improves the confidence in model predictions.

Similar to other findings (Zhang et al., 2022; Dzanic et al., 2020; Schwarz et al., 2021; Tajmirriahi et al., 2022), we confirm there is a considerable shift in both the image and frequency domain between synthetic and real images, which affects the overall realism of such images and may lead to the generation of insufficient surrogates for training. In fact, current state-of-the-art generative adversarial models

used for medical image synthesis fail to transfer low-level and higher frequency components accurately during generation. We demonstrate that even a simple post-processing step, which considers the transfer of frequency components, plays a vital role in generating realistic images and improving the performance of DL-based neural networks trained with synthetic data for segmentation tasks. Similar results are observed when utilizing synthetic images for augmentation.

Our experiments further demonstrate that using synthetic images for training segmentation models requires a significantly larger number of images to achieve similar performance compared to using real images, even with diverse variations in tissue shape and appearance. Similar observations are noted for augmentation purposes, whereby adding a significantly higher number of variable synthetic images is required to achieve statistically significant improvements. This could be attributed to the overall domain shift between synthetic images and real images but also implies that conditional models still produce highly-correlated images with insufficient variation to enhance the network's discriminative ability. Nevertheless, creating these images demands considerably fewer resources than obtaining data through a proper MR acquisition procedure, particularly for tumor data that depend on the limited number of patients available at a particular hospital or site. We also observe that networks augmented with synthetic images particularly struggle with small structures of irregular shapes, as well as areas at the border between different tumor regions (see Fig. B.1 of Appendix B). This is typically a side-effect of blurred edges and low contrast in these regions, commonly found in synthetic images. Despite attempting to address this by considering the low-level Fourier image components, it continues to be a significant source of prediction errors.

Although the proposed work demonstrates improvement, it has some limitations. Firstly, the experiments are conducted on a relatively small dataset, particularly for inference evaluation. Additionally, the dataset only includes HGG tumors, while it would be valuable to investigate the impact of the method on LGG and other tumor types. The overall method should further be tested in a clinical environment to assess the true benefit of both the synthesis and FDA in the clinical practice, as well as validate the generalization of the approach to other segmentation architectures. Finally, while we uncover the potential problems of image synthesis methods in medical imaging and emphasize the importance of considering the spectral characteristics of generated images, further studies are needed to uncover the extent to which both low- and high-frequency components influence the realism of synthetic data, as well as the training and segmentation performance of DL-based networks. By leveraging methods like FDA, we can partially address challenges related to the generation of appropriate spectral image characteristics during synthesis. However, alterations in the Fourier spectra can potentially introduce artifacts in the image domain. In the current approach, we try to mitigate the artifacts by selecting the source and target images based on their structural similarity, which can also be done through visual saliency and spectral similarity indices such as SR-SIM. Careful control of parameters influencing the FDA process is also critical, such as the size of the component swapping window β that requires balancing between smoothing out frequency discrepancies and preserving the original spectral information. However, complete artifact removal may not always be feasible. Future research should focus on methods addressing these discrepancies, potentially combining synthesis in both image and frequency domains for results resembling real clinical images.

Synthetic medical data has the advantage of preserving patient privacy, as it incorporates variations into the original anatomy that eliminate all traceability. However, future research should investigate how much variation conditional GAN models like SPADE GAN introduce into the training data and how much of the original information is retained. To enhance the realism of synthetic images and address discrepancies in imaging and frequency domains, we aim to explore the effectiveness of integrating frequency translation and learning into the standard GAN paradigm. Fourier transform can be utilized in the training process to improve the quality of synthesized images and overcome existing domain shifts. Finally, Fourier methods could be used to measure the quality of generated images and aid in developing strategies to improve the realism of synthesized data used for neural network training.

7. Conclusion

In this work, we study the feasibility of synthetic MR images for the replacement of real MR sequences when training multi-modal brain tumor segmentation networks. We demonstrate that utilizing a conditional GAN for synthesizing multi-modal brain MRI sequences, guided by detailed semantic label maps of both brain and tumor tissue, has a good potential to effectively address such limitations. Through the application of morphological and elastic transformations to the semantic maps of the entire brain and tumor tissue, we can create an extensive synthetic dataset comprising diverse brain tumors. This dataset has the potential to train segmentation models without relying heavily on a large volume of real annotated multi-modal data.

However, we experimentally confirm that there exists a discrepancy between GAN-based synthetic and real images, which is partly due to an existing domain shift between frequency components of two image types. By adapting the low-level amplitude information of synthetic images to those of real images, per MR modality, we obtain images of more realistic appearance and partially tackle the existing domain shift. Such images are not only shown to significantly improve the performance of the model trained with an entire set of synthetic data, but also improve the ability of synthetic images to aid with augmentation. Our findings offer a promising approach to addressing data scarcity in medical imaging segmentation and highlight the importance of considering image frequency in generative approaches for medical image synthesis.

CRediT authorship contribution statement

Yasmina Al Khalil: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Aymen Ayaz:** Conceptualization, Methodology, Software, Validation, Investigation. **Cristian Lorenz:** Writing – review & editing, Supervision. **Jürgen Weese:** Writing – review & editing, Supervision. **Josien Pluim:** Writing – review & editing, Supervision. **Marcel Breeuwer:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yasmina Al Khalil and Aymen Ayaz report financial support was provided by Marie Curie Innovative Training Networks (ITN) fellowship program under project No. 764465. Marcel Breeuwer reports a relationship with Philips Healthcare that includes: employment. Cristian Lorenz and Jürgen Weese report a relationship with Philips GmbH Innovative Technologies that includes: employment.

Data availability

Data will be made available on request.

Acknowledgments

This research is a part of the openGTN project, supported by the European Union in the Marie Curie Innovative Training Networks (ITN), The Netherlands fellowship program under project No. 764465.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2024.102332.

References

AlBadawy, E.A., Saha, A., Mazurowski, M.A., 2018. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Med. Phys. 45 (3), 1150–1158.

Amirrajab, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M., 2022. Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. Comput. Med. Imaging Graph. 101, 102123.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.

Brosch, T., Saalbach, A., 2018. Foveal fully convolutional nets for multi-organ segmentation. In: Medical Imaging 2018: Image Processing. Vol. 10574, International Society for Optics and Photonics, p. 105740U. http://dx.doi.org/10.1117/12.2293528.

Chen, R., Smith-Cohn, M., Cohen, A.L., Colman, H., 2017. Glioma subclassifications and their clinical significance. Neurotherapeutics 14 (2), 284–297.

Chen, Y., Yang, X.-H., Wei, Z., Heidari, A.A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., Guan, Q., 2022. Generative adversarial networks in medical image augmentation: a review. Comput. Biol. Med. 105382.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. 65 (5), 545–563.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp. 424–432.

Dang, K., Vo, T., Ngo, L., Ha, H., 2022. A deep learning framework integrating MRI image preprocessing methods for brain tumor segmentation and classification. IBRO Neurosci. Rep. 13, 523–532.

Durall, R., Keuper, M., Keuper, J., 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7890–7899.

Dzanic, T., Shah, K., Witherden, F., 2020. Fourier spectrum discrepancies in deep network generated images. Adv. Neural Inf. Process. Syst. 33, 3022–3032.

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, 321–331.

Frigo, M., Johnson, S.G., 1998. FFTW: An adaptive software architecture for the FFT. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). Vol. 3, IEEE, pp. 1381–1384.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J.L., Sugano, N., Sato, Y., 2018. Cross-modality image synthesis from unpaired data using CycleGAN. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 31–41.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Işın, A., Direkoğlu, C., Şah, M., 2016. Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput. Sci. 102, 317–324.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Khalil, Y.A., Ayaz, A., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M., 2022. A stratified cascaded approach for brain tumor segmentation with the aid of multi-modal synthetic data. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. Springer, pp. 92–101.

Khayatkhoei, M., Elgammal, A., 2022. Spatial frequency bias in convolutional generative adversarial networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 7152–7159.

Kong, F., Shadden, S.C., 2020. A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 287–296.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H.-K., Jani, A.B., Curran, W.J., Mao, H., Liu, T., Yang, X., 2019. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med. Phys. 46 (8), 3565–3581.

Li, Z., Xia, P., Rui, X., Li, B., 2023. Exploring the effect of high-frequency components in GANs training. ACM Trans. Multimed. Comput. Commun. Appl. 19 (5), 1–22.

Li, J., You, J., Wu, C., Dai, Y., Shi, M., Dong, L., Xu, K., 2018. T1–T2 molecular magnetic resonance imaging of renal carcinoma cells based on nano-contrast agents. Int. J. Nanomedicine 13, 4607.

Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.-A., 2021. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1013–1023.

Magadza, T., Viriri, S., 2021. Deep learning for brain tumor segmentation: a survey of state-of-the-art. J. Imaging 7 (2), 19.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. 3DV, Ieee, pp. 565–571.

Mizutani, R., Saiga, R., Takekoshi, S., Inomoto, C., Nakamura, N., Itokawa, M., Arai, M., Oshima, K., Takeuchi, A., Uesugi, K., et al., 2016. A method for estimating spatial resolution of real image in the Fourier domain. J. Microsc. 261 (1), 57–66.

Mok, T.C., Chung, A., 2018. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In: International MICCAI Brainlesion Workshop. Springer, pp. 70–80.

Myronenko, A., 2019. 3D MRI brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer, pp. 311–320.

Nalepa, J., Marcinkiewicz, M., Kawulok, M., 2019. Data augmentation for brain-tumor segmentation: a review. Front. Comput. Neurosci. 13, 83.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346. http://dx.doi.org/10.1109/CVPR.2019.00244.

Qamar, S., Ahmad, P., Shen, L., 2021. HI-Net: Hyperdense inception 3D UNet for brain tumor segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6. Springer, pp. 50–57.

Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H., Menze, B., 2020. Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. In: Medical Imaging with Deep Learning. PMLR, pp. 655–668.

Schwarz, K., Liao, Y., Geiger, A., 2021. On the frequency bias of generative models. Adv. Neural Inf. Process. Syst. 34, 18126–18136.

Shin, H.-C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M., 2018. 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. Springer, pp. 1–11.

Singh, H., Saini, S.S., Lakshminarayanan, V., 2021. Real or fake? Fourier analysis of generative adversarial network fundus images. In: Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. Vol. 11601, SPIE, pp. 103–109.

Soomro, T.A., Zheng, L., Afifi, A.J., Ali, A., Soomro, S., Yin, M., Gao, J., 2022. Image segmentation for MR brain tumor detection using machine learning: A review. IEEE Rev. Biomed. Eng..

Tajmirriahi, M., Kafieh, R., Amini, Z., Lakshminarayanan, V., 2022. A dual-discriminator Fourier acquisitive GAN for generating retinal optical coherence tomography images. IEEE Trans. Instrum. Meas. 71, 1–8.

Takahashi, S., Takahashi, M., Kinoshita, M., Miyake, M., Kawaguchi, R., Shinojima, N., Mukasa, A., Saito, K., Nagane, M., Otani, R., et al., 2021. Fine-tuning approach for segmentation of gliomas in brain magnetic resonance images with a machine learning method to normalize image differences among facilities. Cancers 13 (6), 1415.

Tiwari, A., Srivastava, S., Pant, M., 2020. Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. Pattern Recognit. Lett. 131, 244–260.

Wadhwa, A., Bhardwaj, A., Verma, V.S., 2019. A review on brain tumor segmentation of MRI images. Magn. Reson. Imaging 61, 247–259.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021. Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 109–119.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.-Y., 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sin. 4 (4), 588–598.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2019. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. Front. Comput. Neurosci. 13, 56.

Wenzel, F., Meyer, C., Stehle, T., Peters, J., Siemonsen, S., Thaler, C., Zagorchev, L., Initiative, A.D.N., et al., 2018. Rapid fully automatic segmentation of subcortical brain structures by shape-constrained surface adaptation. Med. Image Anal. 46, 146–161.

Yang, Y., Lao, D., Sundaramoorthi, G., Soatto, S., 2020. Phase consistent ecological domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9011–9020.

Yang, Y., Soatto, S., 2020. Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4085–4095.

Zhang, L., Li, H., 2012. SR-SIM: A fast and high performance IQA index based on spectral residual. In: 2012 19th IEEE International Conference on Image Processing. IEEE, pp. 1473–1476.

Zhang, Z., Li, Y., Shin, B.-S., 2022. C2-GAN: Content-consistent generative adversarial networks for unsupervised domain adaptation in medical image segmentation. Med. Phys..

Zhao, J., Meng, Z., Wei, L., Sun, C., Zou, Q., Su, R., 2019. Supervised brain tumor segmentation based on gradient and context-sensitive features. Front. Neurosci. 13, 144.

Zhou, T., Ruan, S., Canu, S., 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3, 100004.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.