BORDE: Boundary and Sub-Region Denormalization for Semantic Brain Image Synthesis

Israel N. Chaparro-Cruz[a] and Javier A. Montoya-Zegarra[a,b]
[a] Department of Computer Science, Universidad Católica San Pablo, Arequipa, Perú
[b] Institute for Biomedical Engineering, ETH Zurich, Zurich, Switzerland

***Abstract*—Medical images are often expensive to acquire and offer limited use due to legal issues besides the lack of consistency and availability of image annotations. Thus, the use of medical datasets can be restrictive for training deep learning models. The generation of synthetic images along with their corresponding annotations can therefore aid to solve this issue. In this paper, we propose a novel Generative Adversarial Network (GAN) generator for multimodal semantic image synthesis of brain images based on a novel denormalization block named BOundary and sub-Region DEnormalization (BORDE). The new architecture consists of a decoder generator that allows: (i) an effectively sequential propagation of a-priori semantic information through the generator, (ii) noise injection at different scales to avoid mode-collapse, and (iii) the generation of rich and diverse multimodal synthetic samples along with their contours. Our model generates very realistic and plausible synthetic images that when combined with real data helps to improve the accuracy in brain segmentation tasks. Quantitative and qualitative results on challenging multimodal brain imaging datasets (BraTS 2020 [1] and ISLES 2018 [2]) demonstrate the advantages of our model over existing image-agnostic state-of-the-art techniques, improving segmentation and semantic image synthesis tasks. This allows us to prove the need for more domain-specific techniques in GANs models.**

I. Introduction

According to the World Health Organization (WHO), brain stroke represents the second leading cause of death worldwide happening every 40 seconds and every 4 minutes someone dies from this disease around the world [4]. Besides, among the different types of brain neoplasms, gliomas are the most common with various heterogeneous histological sub-regions characterized by varying intensity profiles [5] (See Fig. 1.b). Because of these reasons, the segmentation of brain tumors in multimodal MRI is one of the most important and challenging tasks in medical applications [1].

Accurate segmentation of brain images like tumors (e.g. gliomas) or strokes (e.g. ischemics) is decisive for diagnosis and treatment. Recent advances in deep learning have shown promising results in the automatic segmentation of brain MRI images [6]. However, most of these approaches require a massive amount of annotated datasets; nonetheless, especially in the medical domain, such annotations are expensive to acquire, require expert annotation level, and can be limited due to privacy issues. Furthermore, due to inter-observer variability, the annotations are not necessarily consistent.

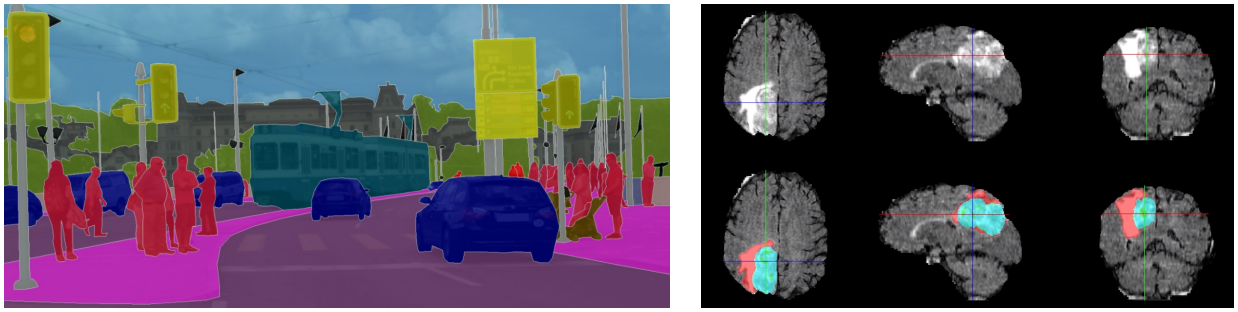
In this context, generative models such as Generative Adversarial Networks (GANs) [7] can be used to generate additional annotated training data. GANs are based on the Nash equilibrium and use a contest between a generator and discriminator to generate highly realistic outputs [8]. Among the first efforts in GANs, Conditional GANs (cGANs) [9] were trained to generate realistic-looking images conditioned on the input data (e.g. class labels, text, images). If the conditional input data is a semantic map, then the model can be used for semantic image synthesis tasks.

The task of semantic image synthesis is still challenging, many state-of-the-art models [10]–[12] use scene (e.g. ADE20K, COCO-stuff, Cityscapes), or face (e.g. CelebAMask-HQ) benchmark datasets with complex scenes and object occlusions. We noticed that those datasets differ quite a lot from medical images datasets:

In scene datasets like Cityscapes [3] (see Fig. 1.a) we have a 2D projection of 3D objects from specific camera coordinates (i.e. a photo), if we vary the camera position we will see a different image; furthermore, each object does not need more than its segmentation mask to be generated.

On the other hand, brain images (e.g. MRI or CT) datasets like BraTS [1] (see Fig. 1.b) are multimodal (e.g. T1, T1c, T2, FLAIR) 2D slices from a brain (3D) in a specific plane (e.g. axial). Ischemic strokes or tumor lesions datasets don't present occlusions, instead, they have sub-regions (e.g. enhancing tumor) of the same object (brain); additionally, due to the type of medical condition (ischemic stroke/blood flow block or tumor/abnormal growth of body tissue), there exists interdependence among brain sub-regions. Moreover, medical image segmentation usually requires additional boundary and organ shape identification [13].

In this paper, our main contributions are: (i) we propose a novel generator architecture, which feeds sequentially a-priori information into the generator based on boundary and brain sub-regions, (ii) we avoid mode-collapse during training by injecting different scales of noise into the model, (iii) besides generating realistic-looking multimodal brain images, our model also generates plausible contours in a multi-task objective, which help to improve the quality of the generated CT and MR images, and (iv) we conduct extensive experimental studies on two public datasets (ISLES 2018 [2] and BraTS 2020 [14] datasets) and show that BORDE outperforms image-



(a) (b)

Fig. 1. Motivation: (a) Street scene of Zurich from Cityscapes dataset [3] is a 2D projection of 3D objects from specific camera coordinates, exists occlusions, and objects are independent; (b) Slice of a flair MRI of a BraTS 2020 dataset [1] in the three possible planes (i.e. axial, sagittal, coronal), we notice that exist sub-regions instead of objects and occur an inter-dependence between these. Scenes and Medical Images are quite different.

agnostic state-of-the-art techniques, such as SPADE [10] or SEAN [11], when used for brain MRI segmentation tasks. This allows us to prove the need for more domain-specific techniques in GANs models.

II. Related Work

Originally, GANs have been proposed as an unsupervised generative framework [7], where random noise variables sampled from known distributions are mapped to realistically looking images [8]. In the ideal case, the data distribution learned from the generator approximates the unknown data distribution.

More recently, with the advent of conditional Generative Adversarial Networks [9] (cGANs), GANs have been turned into supervised generative models by conditioning both the generator and the discriminator with prior knowledge. If the prior knowledge is a semantic map, then we are targeting a semantic image synthesis problem. Indeed, the literature has shown promising approaches for GAN-based semantic image synthesis.

One seminal work is Pix2pix [15], which comprises: (i) an encoder-decoder generator that takes as input a semantic map, and (ii) a PatchGAN-based classifier as a discriminator. Thenceforth, different architectures and loss modifications have been proposed [10], [16], [17] to improve Pix2pix's image synthesis quality.

In [10], the authors noticed that the normalization layers (e.g. Batch Normalization, Instance Normalization) tend to "wash away" the input semantic information. For example, previous methods cause a signal collapse of semantic information when the input consists of a uniform segmentation map (e.g. semantic map is all sky or all grass). To address this issue, the authors propose using the input semantic map to denormalize or modulate the model activations through a normalization layer named SPADE. Their model consists of a generator based on a decoder architecture, which injects semantic information at different scales of the model through the addition of conditional normalization layers that achieves superior performance in semantic image synthesis and allows control over style using a computed embedding vector input.

More specifically, SPADE makes the denormalization part of normalization layers dependent and also spatially-adaptive to the semantic input. Because being a relatively simple but effective semantic image synthesis model, SPADE has been used in recent approaches as a baseline [11], [12], [18].

Normalization layers are basic components of CNN architectures. They help to stabilize the learning process and speed up the training process in two steps. First, normalization layers normalize the network activations into zero mean and unit deviation outputs. Second, the now normalized activations are denormalized with learnable modulation scale/shift parameters. If the scale/shift parameters do not depend on the external input, then the normalization layer is unconditional. Examples of unconditional normalization layers include Batch Normalization [19] and Instance Normalization [20].

If the scale/shift parameters depend on the external input, then the normalization layer is conditional. Conditional normalization layers have been used for style transfer tasks. The most popular ones are Conditional Instance Normalization (Conditional IN) [21] and Adaptive Instance Normalization (AdaIN) [22], where the style information is denormalized in terms of modulation scale/shift parameters.

A more recent approach is SEAN [11]. This model introduces a simple but effective block conditioned on a segmentation mask that describes the semantic regions in the output image. More specifically, SEAN introduces semantic information through normalization layers and style information obtained from a style encoder. The generator architecture of this model is similar StyleGAN [23] but applied to semantic image synthesis.

III. Semantic Image Synthesis

A. Boundary and Sub-Region Denormalization (BORDE)

Compared to existing denormalization approaches, such as SPADE [10] or SEAN [11], we propose a novel denormalization technique intended for brain image synthesis. Given a boundary or semantic mask/map as input, we propose a technique called BOundary and sub-Region DEnormalization (BORDE).

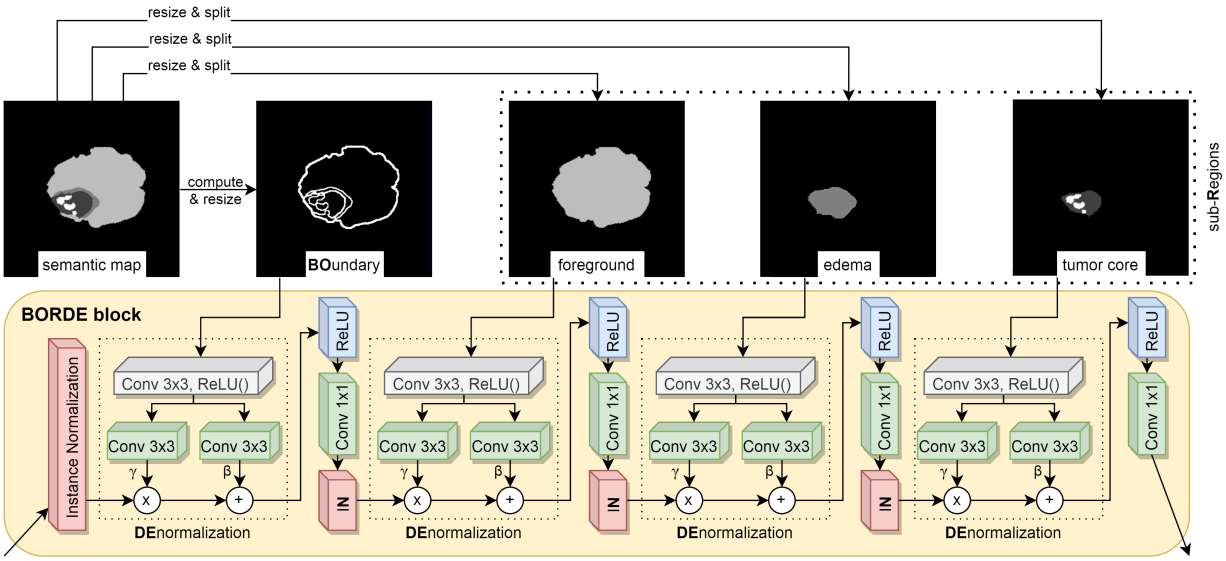


Fig. 2. BORDE block: Boundary mask and a split semantic map are injected through denormalization in an incremental combination to generate realistic-looking brain images.

Mode formally, given the a-priori information $\mathbf{M}$ (boundary mask, foreground or sub-regions mask/map), the goal of the BORDE block is to inject the a-priori information directly on the denormalization layer.

Let C, H, and W denote respectively the number of channels, height, and width of the activation layer. Then, the activation value at site $(c \in C, y \in H, x \in W)$ is given by:

$$\gamma_{c,y,x}(\mathbf{M}) \left(\frac{h_{c,y,x} - \mu_c}{\sigma_c} \right) + \beta_{c,y,x}(\mathbf{M}) \tag{1}$$

where $h_{c,y,x}$ is the activation to be normalized by the mean (μ_c) and the standard deviation (σ_c) in the activations of channel c:

$$\mu_c = \frac{1}{HW} \sum_y \sum_x h_{c,y,x} \tag{2}$$

$$\sigma_c = \sqrt{\frac{1}{HW} \sum_y \sum_x \left((h_{c,y,x})^2 - (\mu_c)^2 \right)} \tag{3}$$

We can see that the denormalization process is given by the calculation of the parameters γ and β and applied to the normalized activation map considering two important characteristics: (i) *spatial-variance* by multiplying and adding respectively the parameters γ and β based on the (y, x) positions. In fact, by making the modulation parameters spatially-invariant (i.e. γ and μ vary only through c) and replacing $\mathbf{M}$ with a real image, we can arrive at a similar formulation as AdaIN [22], (ii) *instance-specific normalization*, since the parameters γ and β are computed per instance. Indeed, by using batch processing we can arrive at the SPADE [10] formulation, (iii) *a-priori semantic segmentation information*. In effect, by replacing $\mathbf{M}$ with style information and by calculating the parameters γ and β per batch, we can arrive at Conditional Batch Normalization [21].

These properties allow improving the image quality, especially in smaller sub-regions, by incrementally injecting a-priori information.

B. BORDE block

Given an a-priori input, such as a boundary or semantic mask/map, we propose a block called BORDE to generate realistic-looking brain images with very fine details characterized by precise contours and localized texture information. An illustration of the proposed block is presented in Figure 2.

In a BORDE block, the input information is split over different normalization layers. Intuitively, all these masks/maps account for a-priori information. First, a binary boundary mask is computed to characterize the shape information of the brain sub-regions. The boundary mask is next fed into a set of convolutional and activation layers to learn the denormalization/modulation parameters to adapt the scales and biases of generator activations. Second, a foreground/background mask used as additional input is injected to guide the attention of the brain style generation. As such, irrelevant background information becomes less relevant. Third, successive semantic masks/maps of sub-regions are used as auxiliary inputs to the BORDE block. This helps to generate semantic-coherent synthetic images.

All the aforementioned input masks/maps are fed into a set of convolutional blocks: First, an intermediate embedding space of 128 is generated through a convolution layer (3×3 kernel, padding 1, and stride 1) followed by a ReLU activation function. Second, two separated convolutions layers (3×3 kernel, padding 1, and stride 1) are applied to learn the denormalization specific-parameters, such as scale/shift. All these convolutional and activation layers that learn the denormalization parameters are replicated for each auxiliary input to the BORDE block.

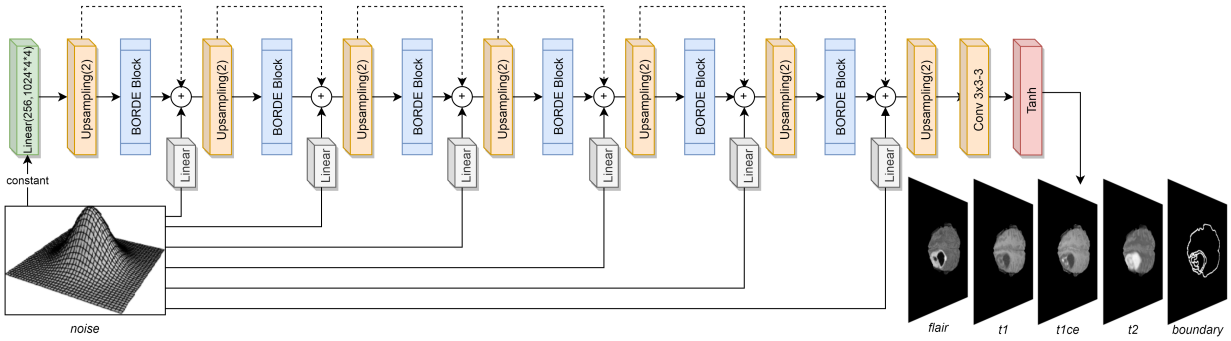


Fig. 3. BORDE generator: it is built upon BORDE blocks and incorporates residual random noise at different levels to (i) generate fine-detailed brain images and (ii) to avoid mode-collapse. Additionally, it is based on residual learning to make the training process more stable and to avoid the gradient vanishing problem.

After every input mask/maps injection, an activation function (ReLU) followed by a convolutional operation (1×1 kernel, stride 1 with zero-padding) is performed. At this point, the number of channels is divided by 2 through the convolution layer in a dimensionality reduction way [24]. Throughout the BORDER block, the size of the activation maps is maintained.

Furthermore, note that each BORDE block operates at different scales. Therefore, we downsample the a-priori information to match the corresponding spatial resolution. Thus, our model is capable of generating realistic-looking multi-resolution brain images.

It is important to note that the structure of the BORDE block is not static and varies according to the number of subregions in the problem. The presented BORDE block in Fig. 2 is based on the BRATS 2020 dataset [1].

C. BORDE Generator

An overview of our generator based on BORDE blocks is given in Figure 3. Our generator relies on a decoder architecture built by BORDE blocks. At each stage, the feature maps are first upsampled ($2\times$) using bilinear interpolation and fed successively into a BORDE block (where semantic and other information are injected). Furthermore, noise is injected into the model in two different ways: (i) a constant Gaussian input noise of size 256 to represent the general sketch of brain images, (ii) different auxiliary inputs in the form of per channel noise scaled vectors. In addition, residual connections are added at each stage to make the training more stable and to avoid gradient vanishing problems [25]. Finally, upon reaching the desired output size, an additional convolution block is used to match the number of output channels followed by an activation function (hyperbolic tangent).

In [26], the authors proposed a generation of intermediate boundary image of the segmentation map, in a coarse-to-fine boundary-aware GAN, to improve brain tumor segmentation. We transform that idea making out generator be multi-task generating images of the required modalities (i.e. T1, T1c, T2, FLAIR) and also an image of the segmentation contours (boundary) where the foreground contour is also included.

IV. Experimental Results

A. Implementation Details

We base our discriminator on the pix2pixHD architecture [15] which concatenates the segmentation map with the image generated as input. Within the network, a convolution (4×4 kernel, 2 stride) reducing the image size by 2 and doubling the number of channels (the output channel size of the first convolution is 64) is performed. Followed by instance normalization [20] and LeakyReLU activation function (slope of 0.2) are performed. These 3 layers are repeated until the size of the output is 16x16, where a 1×1 convolution is finally applied.

We train our model using the Adam optimizer [27] with an initial learning rate of $1e-4$ and set the momentum rate to 0.5. All other parameters are set to the default optimizer values, i.e. $\beta_1 = 0$ and $\beta_2 = 0.999$. Furthermore, to measure the difference between the input images and the reconstructed images, we use the adversarial-hinge loss [10].

BORDE is trained for 100 epochs and it takes about 21 hours for the model to converge. Our algorithm is implemented in PyTorch and is trained on an NVIDIA RTX 2060 SUPER. For all experiments, we consider a batch size of 8 and randomly initialize the network weights.

B. Datasets

We separately train and evaluate our model on two public multimodal brain image segmentation datasets and report quantitative and qualitative results:

1) BraTS 2020: The Multimodal Brain Tumor Segmentation Challenge (BraTS) contains multimodal MRI scans from 369 different patients [1], [14], [28]. For each patient, there are 4 different scans, which include: (i) native (T1), (ii) post-contrast T1-weighted, (iii) T2-weighted, and (iv) T2 Fluid Attenuated Inversion Recovery (FLAIR) modalities. Each MRI scan has been resampled and interpolated into a volume of $240 \times 240 \times 155$ voxels. Furthermore, the dataset contains 293 patients with high-grade glioma (HGG) and 76 patients with low-grade glioma (LGG). All the scans have been manually annotated by experienced radiologists and the segmentation

map contains three sub-regions, which denote either GD-enhancing tumor (ET), the peritumoral edema (ED), and the necrotic and non-enhancing tumor core (NCR/NET).

2) ISLES 2018: The Ischemic Stroke Lesion Segmentation challenge (ISLES) contains 63 different multimodal CT perfusion scans [2], [29]. Each scan contains 4 different perfusion maps (CBV, CBF, Tmax, and MTT), which include: cerebral blood volume (CBV), cerebral blood flow (CBF), time to peak of the residue function (TMax), and mean transit time (MTT). Each scan has a size of 256×256 along the axial dimension with a variable number of slices ranging from 2 to 18. The scans have been manually annotated by radiologists and the segmentation masks contain two sub-regions, which denote either stroke lesions or healthy tissues.

3) Pre-processing: For both datasets, we first convert the 3D volumes and their annotations into stacks of 2D images and add the foreground of the brain as part of the segmentation map. Each 2D semantic map is then fed into BORDE block as input. All 2D slices are resized to a resolution of 256×256 using bilinear interpolation.

C. Baselines

We compare BORDE with 2 leading semantic image synthesis models: SPADE [10] and SEAN [11]; considering SPADE as the current state-of-the-art on semantic image synthesis framework, and SEAN as work with uses style information to improve semantic image synthesis. Furthermore, we compare our results with the best values in the available online leaderboard of both datasets.

D. Evaluation

1) Quantitative results: We train our proposal and baselines with both datasets to generate synthetic images and use them as Data Augmentation (DA) in a segmentation network, namely a U-Net [30] to measure the performance. We use the same training parameters as BORDE to train our in-house developed U-Net and use a cross-entropy loss for pixel-wise label segmentation.

We adopt the same official evaluation protocols provided in the brain segmentation challenges for the quantitative results: For BraTS 2020, we report results on DICE [31] and sensitivity scores. For ISLES 2018, we calculate the DICE, precision, and recall metrics [32].

Table I reports the quantitative results in the semantic segmentation task on the BraTS 2020 dataset. When the generated images from BORDE are used as additional training data for the U-Net, our model improves by at least 7% and 12% in DICE coefficient and the sensitivity respectively compared to the baselines models on Enhancing Tumor (ET) and Tumor Core (TC) regions. Additionally, our model is slightly below SEAN on Whole Tumor (WT) DICE and sensitivity on semantic segmentation task. On average our BORDE model outperforms the results of other models as Data Augmentation in segmentation task on BraTS 2020 dataset.

On the other hand, Table II reports quantitative results on the ISLES 2018 dataset. Our model applied as Data Augmentation on U-Net improving by at least 2%, 2%, and 4% in DICE, recall, and precision respectively. Clearly, our BORDE model outperforms the results of other models as Data Augmentation in segmentation task on ISLES 2018 dataset.

TABLE I
QUANTITATIVE RESULTS IN SEMANTIC SEGMENTATION TASK ON BRATS 2020 DATASET AND METRICS

Method	DICE ↑			Sensitivity ↑		
	ET	WT	TC	ET	WT	TC
wo/DA	0.69	0.85	0.79	0.59	0.88	0.84
w/typical DA	0.73	0.89	0.88	0.69	0.91	0.89
SPADE [10]	0.75	0.89	0.89	0.67	0.90	0.90
SEAN [11]	0.82	**0.92**	0.88	0.77	**0.94**	0.91
BORDE (**ours**)	**0.88**	0.91	**0.91**	**0.87**	0.92	**0.94**
Leaderboard	0.93	0.94	0.95	0.96	0.99	0.96

TABLE II
QUANTITATIVE RESULTS IN SEMANTIC SEGMENTATION TASK ON ISLES 2018 DATASET AND METRICS

Method	DICE ↑	Precision ↑	Recall ↑
wo/DA	0.60	0.63	0.66
w/typical DA	0.68	0.68	0.75
SPADE [10]	0.70	0.69	0.80
SEAN [11]	0.73	0.73	0.81
BORDE (**ours**)	**0.75**	**0.76**	**0.83**
Leaderboard	0.90	0.94	1.00

We can note that our results are closer to the leaderboard in the BraTS 2020 dataset than in ISLES 2018. It's important to consider that the BraTS 2020 challenge has 4 sub-regions (foreground, enhancing tumor, edema, and necrotic/non-enhancing) and ISLES only 2 sub-regions (foreground and lesion) Since the BORDE block varies according to the number of sub-regions in the dataset due to its incremental characteristic. We estimate that the above is decisive for taking more or less advantage so that the results in BraTS 2020 dataset are closer to the leaderboard than in ISLES 2018. However, the results are modestly good considering that we are using a basic U-Net segmentation network

Additionally, we evaluate our proposal and baselines in Frechet Inception Distance (FID) [33] to measure the distribution between the real images and the generated outputs from BORDE and our baselines using 10,000 randomly selected slices on the BraTS 2020 dataset and ensuring the selection of slices with the most number of sub-regions. Table III shows an improvement by at least 19% in FID value in comparison of baseline models, lower scores identify better models.

TABLE III
QUANTITATIVE RESULTS IN FID METRIC ON BRATS 2020 DATASET

Method	FID ↓
SPADE [10]	11.84
SEAN [11]	8.81
BORDE (**ours**)	**7.12**

2) Qualitative results: Aside from the quantitative comparison, we provide qualitative results on the BraTS 2020 dataset.

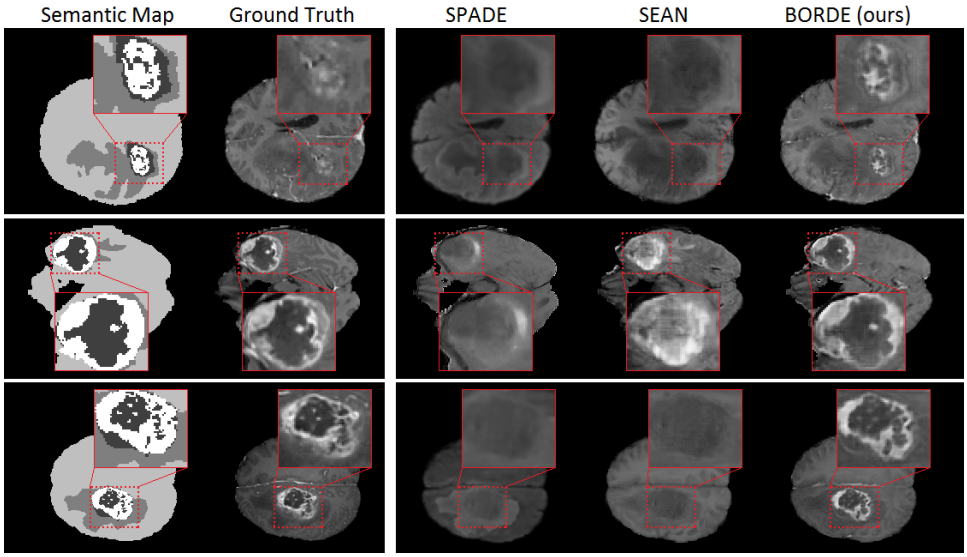


Fig. 4. T1c modality semantic image synthesis of BraTS 2020 dataset. Other models generate artifacts or wash away small sub-regions information. Our model successfully synthesizes all sub-regions details from incremental injection in BORDE blocks.

Fig. 4 shows a qualitative comparison of different semantic image synthesis results of three T1c MRI brain slices. In the case of SPADE, we can see artifacts related to the non-incremental generation of the slide (problem noticed in Section I). Additionally, SEAN does not present artifacts but washes away tumor core information. Finally, BORDE (ours) presents plausible results keeping the information of all sub-regions in the final generated image, with better visual quality and fewer visible artifacts.

Moreover, Fig. 5 shows different visual results of multimodal brain MRI slices together with their corresponding ground truths. Our synthetic images have high image fidelity in all modalities, showing good capabilities in multimodal image generation. A little number of the FLAIR scans in the dataset (shown in rows 3 and 4) were taken in a non-axial plane causing the images to be blurry, this is where we can see that the quality of our generated images is superior.

3) Ablation study: We conduct two ablation studies on the BraTS 2020 dataset and evaluate different configurations of both the BORDE block and the BORDE generator.

Table IV shows possible configurations of the BORDE block. First, it is possible to see a moderate improvement when the boundary is injected along with the semantic map. Second, we can see that the injection of boundary plus sub-regions, in a sequential way, improves significantly the performance over injection of the semantic map in a one-shot manner. Third, we found out that the best performance is on 3 sub-regions (foreground mask, edema mask, tumor-core map) injection over other configurations.

A possible explanation for this behavior is that with 2 sub-regions the incremental process is too short (foreground + lesion) and more similar to a model that injects information non-incrementally without splitting the semantic map. Whilst with 4 sub-regions the process is too large and the incremental injection of information becomes very cumbersome, together with the fact that information from smaller regions (i.e. necrotic and non-enhancing tumor core sub-region) is not present in all the slices. We fixed this BORDE block configuration and use it for quantitative and qualitative experiments, and the next ablation study.

TABLE IV
ABLATION STUDY RESULTS IN BORDE BLOCK ON BRATS 2020 DATASET AND METRICS.

Configuration	DICE ↑			Sensitivity ↑		
	ET	WT	TC	ET	WT	TC
semantic map	0.77	0.89	0.89	0.72	0.89	0.89
boundary + sem. map	0.78	0.89	0.90	0.75	0.89	0.89
boundary + 2 sub-reg.	0.85	0.88	0.89	0.77	0.90	0.90
boundary + 3 sub-reg.	**0.87**	**0.91**	0.91	**0.87**	**0.93**	0.94
boundary + 4 sub-reg.	0.83	0.90	**0.92**	0.84	0.91	**0.95**

TABLE V
ABLATION STUDY RESULTS IN BORDE GENERATOR ON BRATS 2020 DATASET AND METRICS.

Configuration	DICE ↑			Sensitivity ↑		
	ET	WT	TC	ET	WT	TC
BORDE blocks	0.80	0.88	0.90	0.77	0.88	0.89
prev. + multi-task	0.83	0.89	0.89	0.79	0.90	0.92
prev. + residual	0.85	0.87	0.90	0.85	0.91	0.93
prev. + noise	**0.87**	**0.91**	**0.91**	**0.87**	**0.93**	**0.94**

Table V shows the possible configurations of the BORDE generator. First, we show results using only BORDE blocks and generating multimodal outputs. Second, using the *previous* configuration (BORDE blocks) and generating multimodal outputs along with their contours. This can be seen as a multi-task learning process. Third, the *previous* configuration (BORDE blocks + multi-task) and adding residual connections. Fourth, using the *previous* configuration (BORDE blocks

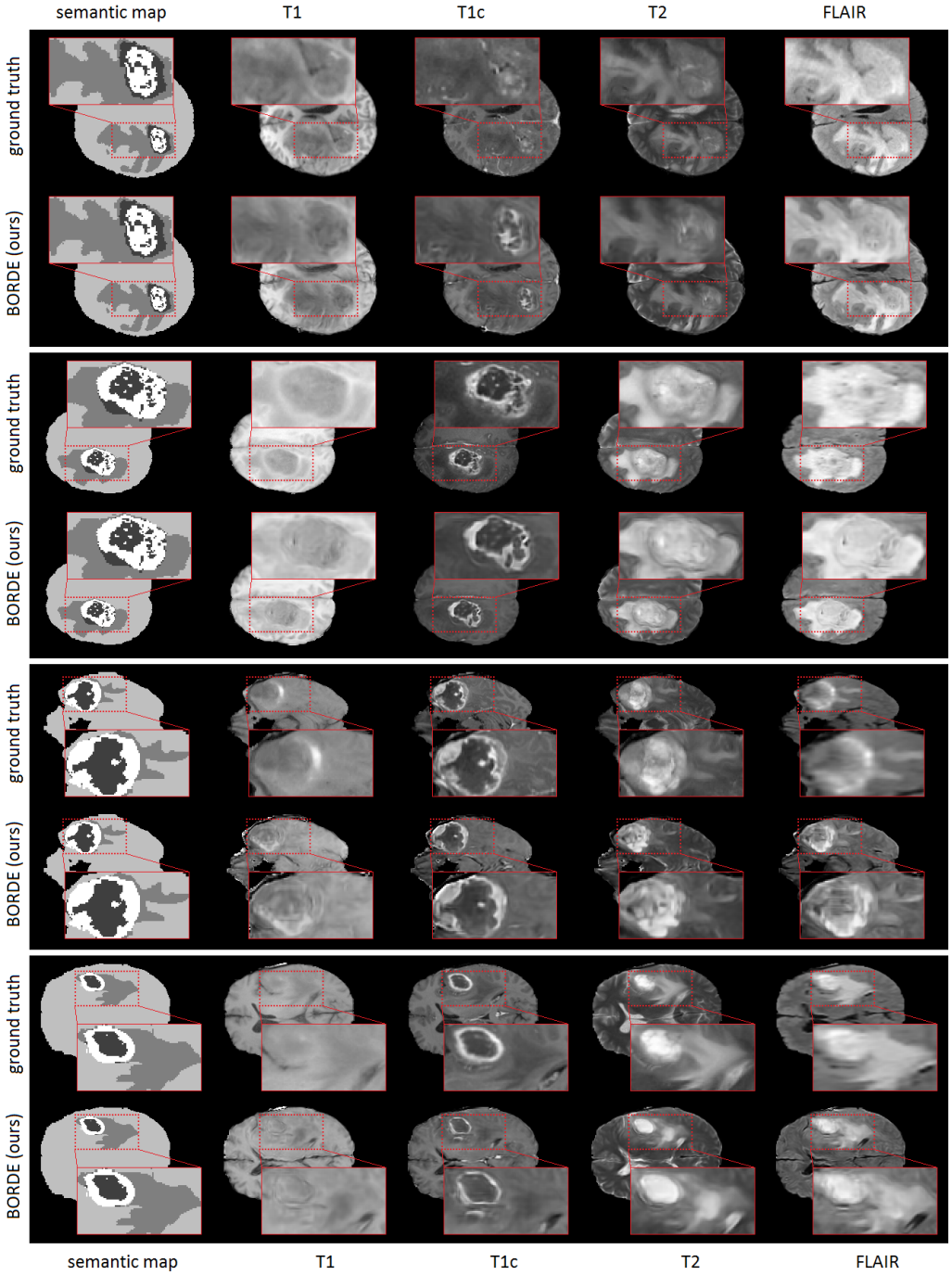


Fig. 5. Our model attains multimodal semantic image synthesis capability and high fidelity. For reference, the ground truth image is shown next to our synthesized images in all cases. It shows that FLAIR ground truth images are blurry due to non-axial MRI reconstruction.

+ multi-task + residual connections) and adding noise as an additional input into each model stage. We can see that each additional component helps to improve the quantitative results. More precisely, the use of a multi-task objective allows having a generator that is more aware of tumor sub-regions. In addition, the injection of multi-stage noise helps to generate more diverse images.

V. Conclusion

We have proposed BORDE, a normalization block, which feeds sequentially a-priori information into the BORDE generator in the form of boundaries and sub-regions through learned denormalization parameters. In addition, we propose a generator, which is formed by BORDE blocks with residual connections, multi-scale noise injection, and multi-task objectives that produce realistic multimodal brain images of tumor

lesions or ischemic strokes. We further demonstrate that our model outperforms different SOTA baselines, such as SPADE or SEAN by decreasing respectively the FID score by 1.69 points. Also, when our synthesized images are used as an additional training source, for semantic image segmentation tasks, BORDE helps to increase the DICE score by 7% and sensitivity by 12%.

ACKNOWLEDGMENTS

This research was supported by the National Fund for Scientific and Technological Development and Innovation (Fondecyt-Perú) within the framework of the "Project of Improvement and Expansion of the Services of the National System of Science, Technology and Technological Innovation" [Grant #028-2019-FONDECYT-BM-INC.INV]. The authors are also grateful to the Swiss National Science Foundation – SNSF (grant # 32003B_159727), a Google Cloud Research Award, and a Titan Xp GPU donation from NVIDIA Corporation. This study was also supported by grant 234-2015-FONDECYT (Master Program) from Cienciactiva of the National Council for Science, Technology and Technological Innovation (CONCYTEC-PERU).

REFERENCES

[1] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[2] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen *et al.*, "Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical image analysis*, vol. 35, pp. 250–269, 2017.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[4] W. Johnson, O. Onuma, M. Owolabi, and S. Sachdev, "Stroke: a global response is needed," *Bulletin of the World Health Organization*, vol. 94, no. 9, p. 634, 2016.

[5] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of mri-based medical image analysis for brain tumor studies," *Physics in Medicine & Biology*, vol. 58, no. 13, p. R97, 2013.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

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

[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[10] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[11] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[12] Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, and N. Yu, "Rethinking spatially-adaptive normalization," *arXiv preprint arXiv:2004.02867*, 2020.

[13] A. Hatamizadeh, D. Terzopoulos, and A. Myronenko, "End-to-end boundary aware networks for medical image segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 187–194.

[14] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "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*, 2018.

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks. corr abs/1611.07004 (2016)," *arXiv preprint arXiv:1611.07004*, 2016.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[17] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8808–8816.

[18] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," *arXiv preprint arXiv:2012.04781*, 2020.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[20] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[21] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.

[22] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[26] T. C. Mok and A. C. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 70–80.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[29] M. Kistler, S. Bonaretti, M. Pfahrer, R. Niklaus, and P. Büchler, "The virtual skeleton database: an open access repository for biomedical research and collaboration," *Journal of medical Internet research*, vol. 15, no. 11, p. e245, 2013.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409

[32] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.