ConText-GAN: using contextual texture information for realistic and controllable medical image synthesis*

Marc-Adrien Hostin, Shahram Attarian, David Bendahan, Marc-Emmanuel Bellemare

ConText-GAN: using contextual texture information for realistic and controllable medical image synthesis*

Marc-Adrien Hostin[1,2], Shahram Attarian[3], David Bendahan [1] and Marc-Emmanuel Bellemare[2]

***Abstract*— This study proposes an enhancement to the ConText-GAN, an image synthesis model using a controllable texture input. The improvement consists in using a texture feature fusion module to reduce the complexity of the model, and enable the use of the OASIS architecture for image generation.**

***Clinical relevance*— The ConText-GAN can be used to generate images of fake patients, which are useful in the medical field due to the scarcity of data. An example is given of the generation of images showing pathological muscle tissue in the context of neuromuscular diseases.**

I. INTRODUCTION

Image synthesis can have multiple applications in medical imaging, such as data augmentation, fake patient synthesis, or multimodal image synthesis [1]. Deep learning methods have made it possible to render synthetic images extremely realistic, but the controllability of these methods remains a major challenge [2].

Controllability is the ability to predict the result of a model according to its input. To be able to predict the result produced by a model, the input information must be sufficient to describe the desired synthetic image. Classically, controlled synthesis methods use label maps, given in the form of segmentation masks. For example, muscle segmentation masks given as input to a generative model have been used to generate lower limb MRI scans [3].

Label maps provide mainly geometric information about the different parts of the image to be generated, such as position and shape. We suggest that adding texture features would improve the controllability of image synthesis models. For example, in addition to generating lower limb muscles localized by a segmentation mask, we could control the texture of the muscular tissue.

A popular method to achieve controllability is to use disentanglement methods [4]. However, it only allows generating samples from a limited number of classes (*e.g.*, red bird, blue car), while we propose to control synthesis using a continuous feature space. In label-to-translation medical applications, a texture patch has been tested as model input condition [5]. This limits the generation of textures to those existing in the patches given as input. Furthermore, it is difficult to assess whether the textures generated are indeed the same as those in the input patch.

To improve texture control, we suggest using statistical texture characteristics, such as entropy, applied to region maps. A region map would contain the values of the desired texture features in that region. For each texture feature to be controlled, a region map is created. We call context map the set of these region maps, given as input to the synthesis model. The complete model, inspired by Generative Adversarial Model (GAN) methods, is called ConText-GAN, which stands for Controllable Texture GAN.

Originally [3], the ConText-GAN architecture was based on pix2pix [7], the benchmark for label-to-image translation models. However, more recent architectures for creating realistic images, such as OASIS, inspired by SPADE [8] and pix2pix, have been developed. The OASIS [6] model uses the SPADE architecture for the generator, and employs a segmentation network, U-Net, as the discriminator. While more powerful than its predecessors, the OASIS model is more complex and pricier to train. Moreover, the addition of feature maps instead of a simple label map further increases the model's complexity. To reduce overall complexity, and enable the use of OASIS, we propose a feature map fusion module. The module takes a set of feature maps as input, learns the best combination and creates a single map that summarizes the needed information, the context map. This stage not only reduces the size of the OASIS input data, but also enables the network to learn which input texture features are most relevant for generating a realistic textured image. This last point will improve interpretability, as we will have access to the weights that the network has assigned to each texture map if needed.

To test our approach, we propose to use the same application as for the first ConText-GAN, *i.e.*, the generation of MRI images of the lower limbs of patients with neuromuscular diseases. To the best of our knowledge, in this field, the only method of data generation by deep learning uses a cycle GAN [9], which does not allow the desired textures to be specified to the model, as we propose in this paper.

We propose to use standard image quality metrics to measure the gains in realism and controllability achieved by using OASIS compared with the previous Pix2Pix-based version. We selected typical image quality indicators, L1 distance, Structural Similarity Index Measure (SSIM), Peak Signal to Noise Ratio (PSNR), Frechet Inception Distance (FID), and a custom metric for controllability. In addition, we demonstrate the gains in accuracy, complexity, and interpretability brought by the fusion module.

*This work was not supported by any organization

[1]Aix Marseille University, CNRS, CRMBM, Marseille, France

[2]Aix Marseille University, CNRS, LIS, Marseille, France

[3]Neurology Department, APHM, Reference Center for Neuromuscular Diseases and ALS, La Timone University Hospital, Aix-Marseille University, Marseille

II. METHODS

A. Architecture

The aim of ConText-GAN is to generate images in response to commands for shapes and local texture features. These commands are embodied in the context map. Let H and W be the height and width of the image, and N_L the number of regions to control. Each region has its shape, and all the regions together form a map that provides the shape of the image components, the image segmentation mask. Let N_F be the number of scalar features that characterize a specific texture. To control the texture within each region, N_F feature maps are created. Each feature map defines the spatial distribution of a texture feature, which is the label value within each region of the segmentation mask.

The process of training ConText-GAN with OASIS is illustrated in Figure 1. First, the feature maps are given to the fusion module, which returns a fused version of the texture features, the context map. The context map is passed to the generator, to create a prediction image ($\hat{y}$). $\hat{y}$ is compared to the reference image (y) using the $\mathscr{L}_G$ function, then fed to the discriminator.

In the case of OASIS, the discriminator is a U-Net network responsible for segmenting $\hat{y}$. The segmentation result ($\hat{s}$) is compared to the reference (s) by the $\mathscr{L}_{GAN}$ function, to check that $\hat{y}$ can be segmented as accurately as a real image. In other words, the discriminator verifies that the label position information is properly encoded in $\hat{y}$.

In the case of Pix2Pix, the discriminator is simply replaced by a PatchGAN classification network, responsible for determining whether $\hat{y}$ is a real or synthesized image, as described in [3].

In both cases, the generator's role is to create an image $\hat{y}$ closest to y to fool the discriminator. The image is considered realistic when the discriminator can no longer distinguish the real image from the synthetic one.

B. The fusion module

The fusion module takes as input the feature maps. The first section of the module is a channel attention component, inspired by CBAM [11], as illustrated in Figure 1. The channel attention associates a weight with each input feature map, the value of which is a learning parameter for the model. In this way, the model can select those that will best optimize its learning. The rest of the module consists of a series of 3D convolutions (with channel size $N_L/64/128/64/1$) that create a linear combination of the feature maps to create the context map.

Finally, the fusion module will have selected the best texture features, and reduced the context map dimension to (N_L, H, W), instead of (N_F, N_L, H, W). The OASIS input data has therefore been changed from 3D+C to 2D+C, where C is the channel dimension. This change makes it possible to use the OASIS generator with 2D convolutions rather than 3D. Assuming that a $2D$ convolution with a $3x3$ kernel contains 9 of parameters, and that a $3D$ convolution with a $3x3x3$ kernel contains 27, we can assume that the complexity of the network would be, at the very least, multiplied by 3. Since the number of parameters in the network is $69M$, this would then be $69M \times 3 = 207M$, requiring 3 times more computing resources. In contrast, the fusion module only adds four 3D convolution layers and contains $450k$ trainable parameters (0.6% of the OASIS number of parameters).

C. Loss functions

The network is trained using two functions: $\mathscr{L}_G$, associated with the generator, and $\mathscr{L}_{GAN}$, associated with the discriminator.

$\mathscr{L}_{GAN}$ is dependent on the type of discriminator. For pix2pix, $\mathscr{L}_{GAN}$ is a cross-entropy function [7]. For OASIS, as the discriminator is a segmentation network, the function is a cross-entropy evaluating per pixel segmentation accuracy [6]. These functions have not been modified in our study, as we have concentrated on the generator, responsible for synthesizing the textures.

In standard pix2pix, $\mathscr{L}_G$ is a L1 distance loss, and OASIS does not use $\mathscr{L}_G$. In the ConText-GAN, two terms were added to $\mathscr{L}_G$. The first term ($\mathscr{L}_S$) is based on SSIM [1] and aims to improve the realism of computer-generated images, by considering the similarity of structural information between predicted and reference images. The second term ($\mathscr{L}_{CT}$) is a texture control term. Given i, j pixel coordinates, $\Theta_{i,j} = (\Theta_{i,j}^{f_0,l_0}, ..., \Theta_{i,j}^{f_{N_F},l_{N_L}})$, with $\Theta_{i,j}^{f_k,l_n}$ where f_k is the kth texture feature, and l_n the n^{th} label region. For instance, f_k could be the entropy measured in a region l_n. $\mathscr{L}_{CT}$ consists in measuring the distance between the texture features measured on the synthetic image ($\Theta(\hat{y})$) and the input context map ($\Theta(y)$) (1). Finally, the whole function $\mathscr{L}_{tot}$ is a weighted sum of the three terms (2). We chose to set $\lambda_1 = \lambda_2 = \frac{1}{2}$.

$$\mathscr{L}_{CT}(y,\hat{y}) = |\Theta(y) - \Theta(\hat{y})| \quad (1)$$

$$\mathscr{L}_{tot}(y,\hat{y}) = \mathscr{L}_G(y,\hat{y}) + \lambda_1 \mathscr{L}_{CT}(y,\hat{y},m) + \lambda_2 \mathscr{L}_S(y,\hat{y}) \quad (2)$$

D. Dataset

We tested our texture generation tool on MRI scans of the lower limbs (thigh and calf in two separated acquisitions). The database contains acquisitions from 14 controls and 88 patients with various neuromuscular diseases. As muscle tissue texture is dependent on the type and extent of disease, texture synthesis is an interesting application in this context. In addition, pathological tissue can hinder automatic segmentation methods [10], which makes the generation of pathological images interesting for data augmentation.

Pathological muscle tissue is represented by the presence of fatty infiltration (Figure 2). We have chosen two texture features relating to infiltration: mean intensity, which represents the amount of white fat in the muscle; entropy, which represents the heterogeneity of the infiltration.

The dataset is a private one from the La Timone Hospital, Marseille. MRI scans were recorded at 1.5T (MAGNETOM

[1] https://github.com/VainF/pytorch-msSSIM

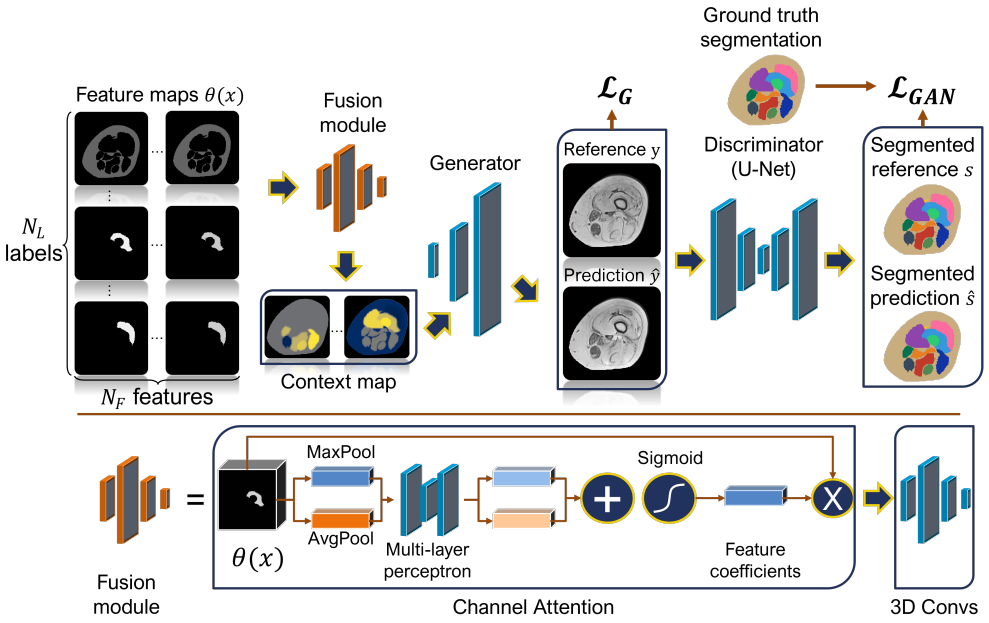

Fig. 1. ConText-GAN training process with OASIS architecture. The feature map $\Theta(x)$ is transformed into a context map by the fusion module, then into an image by the generator. The result is compared to the reference ($\mathscr{L}_G$) and evaluated by the discriminator ($\mathscr{L}_GAN$) to train the model.

Avanto, Siemens Healthineers, Erlangen, Germany) at the thigh level using a spine coil on the bottom and a flexible coil on top of the lower limb. One image set consisted of 2D T_1-weighted MRI (T_{1W}) (TR = 578 ms; TE = 11 ms; FA = 90°; bandwidth = 182 Hz/pixel; in-plane matrix size/voxel size = $320 \times 160/1.26 \times 1.26$ mm^2; 38 slices [slice thickness = 4.40 mm]; slice gap = 0.40 mm). The second set T_{1w} MRI acquisition featured : TR = 549 ms; TE = 11 ms; FA = 120°; bandwidth = 195 Hz/pixel; in-plane matrix size/voxel size = $320 \times 320/0.68 \times 0.68$ mm^2; 20 slices [slice thickness = 10.00 mm]; slice gap = 5.00 mm. For each set, thigh and calves sections were acquired separately (using the same settings). To get a consistent set of images, the images were all resized to 256×256, and a Gaussian blur was applied to the highest resolution images to keep the resolution the same throughout the set.

E. Evaluation

Two criteria were evaluated to determine the performance of our generative networks.

The first is the realism of the synthetic images. To achieve this, we created context maps in which the shape of the regions was deformed by random elastic deformation. These deformed maps were then fed to the network, and the generation results ($\hat{y}$) were compared with reference images deformed with the same displacement field ($E(y)$). $E(y)$ and $\hat{y}$ were compared with standard similarity metrics: L1 distance; PSNR and SSIM. In addition, we created a set of synthetic images generated from deformed regions and randomly sampled context $\hat{Y}_R$. With this entirely new dataset of fake images, we were able to measure the FID, a measure of the distance between the distribution of reference images, and $\hat{Y}_R$.

The second criterion is model controllability. The aim is to know whether the texture command given to the model is respected. To accomplish this, we simply used the $\hat{Y}_R$ set to measure the L1 distance between the texture features present in the synthetic images and the feature maps used to generate them. In our case, this boils down to measuring the mean and entropy over each area of the generated image, and calculating the absolute difference between these values and the mean and entropy values given in the input command. To simplify analysis of the results, we normalized the features between 0 and 100, and averaged the L1 distances of the mean and entropy, resulting in a single controllability score, equivalent to a percentage.

Each criterion was tested on three versions of ConText-GAN based on: pix2pix without fusion module (P2P); pix2pix with fusion module (P2PF); OASIS with fusion module (OASIS).

III. RESULTS

A. Image quality

TABLE I

IMAGE SYNTHESIS REALISM MEASURED FOR P2P, P2PF AND OASIS, ON CALF AND THIGH.

		L1	PSNR	SSIM	FID
Calf	OASIS	0.024 ± 0.011	**22.065 ± 1.646**	**0.919 ± 0.031**	16.691
	P2PF	0.023 ± 0.011	21.208 ± 1.499	0.903 ± 0.039	**15.577**
	P2P	**0.022 ± 0.009**	21.079 ± 1.473	0.899 ± 0.04	22.579
Thigh	OASIS	0.050 ± 0.024	**22.156 ± 1.709**	**0.856 ± 0.057**	**14.462**
	P2PF	**0.042 ± 0.024**	20.76 ± 2.053	0.832 ± 0.072	20.949
	P2P	0.044 ± 0.023	20.938 ± 1.669	0.822 ± 0.069	28.056

For all the networks, the L1 distance between the synthetic and real images is low, since it doesn't exceed 0.05, which corresponds to a 5% error (Table I). As the difference

between the L1 scores of each model is small, it does not allow us to compare the models.

PSNR and SSIM show that the quality of model synthesis follows the order OASIS $>$ P2PF $>$ P2P (except for PSNR on the thigh), with a high average SSIM for the thigh at 0.86 ± 0.06, and for the calf at 0.92 ± 0.03.

FID is clearly improved by the fusion module (Table I), and is better on the thigh for OASIS, and better for P2PF on the calf (although close to OASIS).

The L1 and PSNR metrics, although standard, are pixel-to-pixel distance metrics, and therefore sensitive to the slightest translation. Disregarding these metrics, OASIS shows the best results on the thigh and the best SSIM on the leg, with an acceptable FID. In addition, the fusion module helped to improve the realism of the synthesis, since P2PF had better results than P2P.

B. Texture controllability

TABLE II

CONTEXT-GAN CONTROLLABILITY MEASURED FOR P2P, P2PF AND OASIS, ON CALF AND THIGH.

	OASIS	P2PF	P2P
Calf	5.467 ± 3.815	6.102 ± 6.029	$\mathbf{4.533 \pm 4.435}$
Thigh	5.920 ± 4.176	8.433 ± 8.046	$\mathbf{5.500 \pm 5.835}$

The controllability study indicates that P2P is the most controllable network, although the values for each model are close (Table II). This suggests that the fusion module slightly alters the controllability of the model. We assume that in creating the context map, the fusion module loses some information contained in the feature maps. However, the fusion module still reduced the complexity, and enabled more realistic images to be synthesized, as shown in section III-A.

C. Qualitative analysis

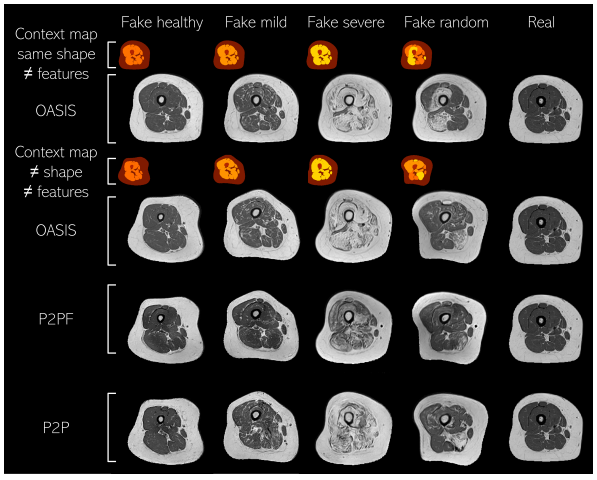

Fig. 2. Qualitative demonstration of the texture synthesis capability of ConText-GAN, with input maps of different shapes and features, and comparison of P2P, P2F and OASIS results on thigh images.

Figure 2 shows the generative capability of ConText-GAN. Using the same label shape, the first line shows that it is possible to generate healthy, weakly and strongly affected muscles. The real image corresponds to the image from which the region map is derived, thanks to a segmentation step. The other three rows show the same results by changing the shape of the regions in each column, for each P2P, P2F and OASIS model. The overall quality of the images confirms the quantitative analysis. P2P images appear to be the noisiest, while P2PF images seem to be the least controlled. Overall, the OASIS-based ConText-GAN provided the most realistic images.

IV. CONCLUSIONS

ConText-GAN allows fine-grained explicit control of MRI generation. With the OASIS architecture, we have proposed an improved version of our first ConText-GAN based on Pix2Pix. The improvement is twofold. First, the image quality benefits from the efficiency of OASIS. Second, the fusion module significantly reduces model complexity and learns how best to combine texture feature maps. The fusion module allows any number of descriptive features to be linearly combined. For simplicity, we have used mean and entropy as texture features, but a later study will explore the effect of using a larger number of features. The improved ConText-GAN was thus able to synthesize realistic MRIs of lower limbs with healthy and pathological muscle textures in a localized and controlled manner. Future applications could involve data augmentation to enrich classification or segmentation training databases.

REFERENCES

[1] Skandarani, Y., Jodoin, P. M., and Lalande, A. (2023). Gans for medical image synthesis: An empirical study. Journal of Imaging, 9(3), 69.

[2] Baraheem, Samah Saeed, Trung-Nghia Le, and Tam V. Nguyen. "Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook." Artificial Intelligence Review (2023): 1-53.

[3] Hostin, M.-A. and Sivtsov, V. and Bendahan, D. and Bellemare, M.-E. ConText-GAN: Controllable context image generation using GANs. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE, 2023. [Accepted, not published yet.]

[4] Li, Yuheng, et al. "Mixnmatch: Multifactor disentanglement and encoding for conditional image generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[5] Oliveira, Dario Augusto Borges. "Controllable skin lesion synthesis using texture patches, bézier curves and conditional gans." 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.

[6] Sushko, V., Schönfeld, E., Zhang, D. et al. OASIS: Only Adversarial Supervision for Semantic Image Synthesis. Int J Comput Vis 130, 2903–2923 (2022).

[7] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[9] Gadermayr, Michael, et al. "Image-to-image translation for simplified MRI muscle segmentation." Frontiers in Radiology 1 (2021): 664444.

[10] Hostin, Marc-Adrien, et al. "The impact of fatty infiltration on MRI segmentation of lower limb muscles in neuromuscular diseases: A comparative study of deep learning approaches." Journal of Magnetic Resonance Imaging (2023).

[11] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.