MM-GAN: 3D MRI Data Augmentation for Medical Image Segmentation via Generative Adversarial Networks

Yi Sun
School of Computer Science
Fudan University
Shanghai, 210043 China
ysun@fudan.edu.cn

Peisen Yuan
College of Information Science and Technology
Nanjing Agricultural University
Nanjing 210095, Jiangsu China
peiseny@njau.edu.cn

Yuming Sun
School of Data Science
Fudan University
Shanghai, 210043 China
18210980013@fudan.edu.cn

***Abstract*—Due to the limited amount of the labelled dataset, which hampers the training of deep architecture in medical imaging. The data augmentation is an effective way to extend the training dataset for medical image processing. However, subjective intervention is inevitable during this process, not only in the pertinent augmentation but also the non-pertinent augmentation. In this paper, to simulate the distribution of real data and sample new data from the distribution of limited data to populate the training set, we propose a generative adversarial network based architecture for the MRI augmentation and segmentation (*MM-GAN*), which can translate the label maps to 3D MR images without worrying about violating the pathology. Through a series of experiments of the tumor segmentation on BRATS17 dataset, we validate the effectiveness of *MM-GAN* in data augmentation and anonymization. Our approach improves the dice scores of the whole tumor and the tumor core by 0.17 and 0.16 respectively. With our method, only 29 samples are used for fine-tuning the model trained with the pure fake data and achieve comparable performance to the real data, which demonstrates the ability for the patient privacy protection. Furthermore, to verify the expandability of *MM-GAN* model, the dataset *LIVER100* is collected. Experiment results on the *LIVER100* illustrate similar outcome as on BRATS17, which validates the performance of our model.**

***Keywords*-MRI Data Augmentation, Medical Image Segmentation, Generative Adversarial Networks**

I. Introduction

The usage of deep learning in medical imaging analysis has already shown a great prospect, including detection [1], lesion diagnosis [2], anatomy registration [3], organ images reconstruction [4], diseases report [5]. Despite that progress, the greatest challenge hindering the development of deep learning in the medical image domain, is the limited amount of annotated samples. So far, the largest CT image dataset *DeepLesion* [6] contains 32,120 CT slices of 4,427 unique patients, and the largest Chest X-ray dataset – *ChestXray14* [7] comprises 108,948 X-ray images of 32,717 unique patients, which magnitude is much less than the *ImageNet* and *MS COCO* dataset of millions of pictures. This is attributable to the unaffordable cost in money and time for experts to accurately delineate the pixel-wise boundary of the different regions and the difficulty of balancing the inconsistencies labels from different annotators.

To alleviate the lack of labeled data, data augmentation is a conventional method in medical image segmentation. For example, non-parametric non-uniform intensity normalization (N3) algorithm [8] is a popular method usable in segmentation task, in order to correct the bias field introduced by the various data source (*i.e*. data acquired by different institutes, scanners, and protocols). Typically, data augmentation techniques can be divided into two categories: pertinent augmentation, and non-pertinent augmentation. Pertinent augmentation, such as registration and normalization, stresses the important features of user expectation and reduce the disturbance of useless feature (*i.e*. the shape and intensity of lesions etc.). Conversely, non-pertinent augmentation actively imposes the disturbance on the original data to improve the robustness of the model. For example, random elastic deformations [9] is a popular procedure in medical image segmentation. However, either pertinent augmentation or non-pertinent augmentation is subjective, which requires users to pre-define the transformations applied to raw data. The most objective and reasonable, but not a realistic way is sampling more data from the real data distribution.

Although sampling more data is expensive and modeling the real data distribution is impossible, we can simulate the distribution of real data by Generative Adversarial Networks (GANs) [10] and produce abundant training examples by this learned model. GANs can convert the discrete distribution of limited training samples into a continuous distribution through a min-max two-player game, thereby provide more variant examples. There are lots of work [11]–[13] successfully improve the model performance by mixing up the real data and fake data produced by GANs. But a large part of these works is focused on 2D images. However, data augmentation using full 3D MR images is more challenging because the pathology

The study was funded by National Natural Science Foundation of China (No.61502236), Shanghai Science and Technology Committee (No.17511104200, 18411952100 and 17411953500), and College Students Entrepreneurship Training Program 2019(No. S20190025).

978-1-7281-8156-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICBK50248.2020.00041

as different MR sequences represents pathology differently.

In this paper, we propose and an end-to-end architecture that can synthesize the 3D MR images by a conditional GAN model, called *MM-GAN*. Users can input the label map modified on an existing label map (*i.e*. tumor's size, shape, and location), and gain a pair including the input label map and the output MRI image through the network. In subsequent experiments, we have done extensive experiments to validate the effectiveness of a larger training dataset populated by synthetic data. The results demonstrate that these pairs significantly benefit the performance of brain tumor segmentation networks. Besides, personal health information has been protected well, which means that the synthetic MRI images can be shared outside of the institution which collected the raw data [12]. The synthetic MRI images show notable variations with raw data that share the same label map (Figure 6, row 2 and 3), and its identifiable information, such as DICOM metadata and skull, was removed. Therefore, identifying the patient by the MRI image is nearly impossible. Furthermore, an origin MRI dataset *LIVER100* is introduced to apply data augmentation and liver lesion segmentation. We implement experiments on *LIVER100* similarly and acquire grateful results. As far in our knowledge, it is the first GAN-based augmentation method in MR imaging of the liver.

The primary contributions of this paper can be divided into two parts:

1) We collected and introduced a new dataset *LIVER100*, which are based on MRI in CT and PET modalities from 100 patients.
2) We proposed an end-to-end framework, called MM-GAN, to synthesize the 3D MR images for medical image segmentation by a conditional GAN model.
3) We verified the effectiveness of the synthetic data generated by *MM-GAN* for data augmentation and anonymization through extensive experiments.

II. Related Work

GANs are increasingly gaining traction in both computer vision and the medical community and has been applied to many domains. Through a minimax two-player game, the generator in GANs will imitate the real data distribution by discriminator criticism, and realize the applications like image translation [14], image synthesis [15], data augmentation [16], [17], image completion [18]. Despite these successful work, the instability of GANs is still a large challenge , which will be more severe in synthesizing high-resolution images [15] or 3D voxels [19]. To address that challenge, there are plenty of regularization techniques been proposed. Takeru Miyato *et. al.* [20] have experimented these methods, including weight clipping, WGAN-GP [21], batch-normalization (BN), layer normalization (LN) , weight normalization (WN) [22], orthogonal regularization [23] and spectral normalization [20] , on CIFAR10 and STL-10 using *inception score* (IS) and *Fréchet inception distance* (FID) . Their results show that orthonormal regularization and spectral normalization outstanding than the other regularization techniques for a variety of settings in training, because these methods avoid the rank deficiencies. Another important phenomenon should be noticed is the performance of orthonormal regularization will be deteriorated when the feature dimensions increase, but spectral normalization would not be affected.

In the community of medical image analysis, GANs have also gained wide attention in assisting radiologists processing medical image data. Benefitted from the ability of simulation and synthesis of GANs, such promising applications as low dose CT denoising [24], image segmentation [25], [26], reconstruction [27], image registration [28], [29] have been introduced. In this paper, we focus on the segmentation task. A common practice [30], [31] in this domain is populating the training data with synthetic data produced by GANs to overcome insufficient labeled data. And there is a large proportion of work constrained by limited hardware devices and the time-consuming training process treated the 3D voxels (i.e. MRI images) as a sequence of 2D slices to synthesize. Shin *et. al.* [12] proposed a method which employs three GANs networks to directly generate 3D synthetic abnormal MRI images with brain tumors. Three GANs are trained for the different image translation tasks, that is, MRI-to-brain segmentation, label-to-MRI synthesis, and MRI-to-tumor segmentation. By manipulating the tumor label map adding on the label map generated by the first GANs (MRI-to-brain segmentation), users can control the location and size of tumors.

Compared with the above pipeline, we conduct the same task in one step by only one generative adversarial network without the tedious translation.

III. The *LIVER100* Dataset

To facilitate the study we need more labeled medical imaging datasets of different organs. In this section, we introduce a new dataset *LIVER100* to extend our data augmentation and segmentation model to liver MR images. We construct the *LIVER100* dataset based on MRI in CT and PET modalities from 100 patients. The slices in CT modalities which cover the liver are manually picked out, and then the focal area is labeled with regular interval by experienced clinicians. In specific, the manually labeled slices take about one-third in the amount of all picked slices. The labels for the others are calculated through linear interpolation. When generating the gold standard, the MR images are converted into NIFTI format, and the labeled focal areas are padded. Furthermore, we have corrected some misoperations in the labels which can guarantee less data noise and better machine comprehension.

The voxel images in *LIVER100* contain about 100 slices on average. Figure 1 gives an example from different views and at different heights. Each segmentation file requires about 25MB of memory each, while the volume file needs about 100MB. The slices with their labels are shown in 1 for instance.

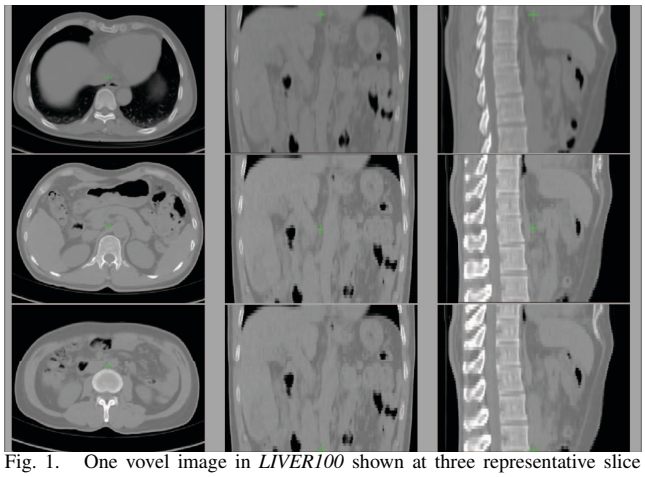

Fig. 1. One vovel image in *LIVER100* shown at three representative slice heights. The three columns show different views of the example.

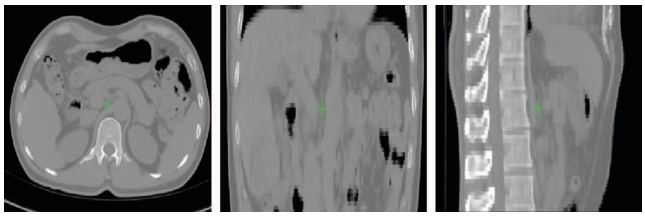

Fig. 2. Label maps in one voxel image in *LIVER100*. The lines in orange draw the outlines of the focal zones in the liver. The generated labels between manually labeled slices have strong anatomical consistency.

IV. Methodology

A. Generator G

In *MM-GAN*, we employ a 3D U-Net [32] as generator G. As in Figure 3, we progressively downsample the input and double its channels simultaneously, through the downsampling block. This process begins to reverse from the bottleneck layer. As long as the output of the upsampling block has equal resolution with the input, this process end. And a $1 \times 1 \times 1$ convolution layer reduces the number of output channels to 1, which is the channel number of MR images. In addition, shortcut connections are added between the two blocks that output the same size feature map, to transmit the high-resolution features.

Different from the 3D U-Net [32], the batch normalization layers [33] in downsampling blocks and upsampling blocks are replaced by instance normalization layers [34] due to its superior performance in minibatch cases (setting to 1 in our case) and generative tasks. The activation function in downsampling and upsampling blocks is LeakyReLU, instead of ReLU. The activation function in the output block is set to Tanh for the generating task. Beyond that, we add a spectral normalization layer [20] after each convolution layer to stabilize the training. The same scheme also presents in Discriminator [35]. Spectral normalization limits the Lipschitz constant of networks by constraining the spectral norm of the weight matrices of each layer. It also has the advantage

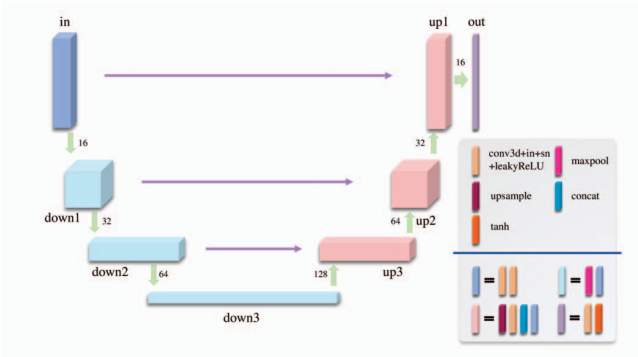

Fig. 3. The architecture of Generator. The generator consists of 4 parts: 1 input block (marked in dark blue), 3 upsampling blocks (marked in bright blue), 3 downsampling blocks (marked in pink) and 1 output block (marked in purple). The internal structures of the 4 parts are shown in the right bottom. It should be noted that the first layer of upsampling blocks halved the channels of input before the concatenate operation.

of not requiring additional hyperparameter tuning and small computational costs.

Practically, we set both the number of downsampling blocks and the number of upsampling blocks to 3, and the number of channels in the first downsampling block to 16, which achieves a balance between GPU memory constraints and computational cost. Meanwhile, we adopt the $3 \times 3 \times 3$ convolutions with stride 1 and $2 \times 2 \times 2$ max-pooling with stride 2. To eliminate the checkerboard artifacts caused by transposed convolution, we replace it with the cubic interpolation upsampling followed by a $3 \times 3 \times 3$ convolution.

B. Discriminator D

SimGAN [36] has proven to be a reliable effective method to stimulate the local features in 2D images by its discriminator. The D classifies all local image patches separately by a limited receptive field, which not only decreases the parameters of D, but also enriched the training samples.

In our task, the local feature should be paid more attention than global features because the brain scans data used in our experiments are co-registered on the same anatomical template and the main structure of brains is similar. Thereby, the above mechanism benefits our task perfectly. Here, we extend their discriminator to 3D cases. Figure 4 shows the architecture of our discriminator. The last layer of D outputs a $w \times h \times d$ dimensional probability map of 3D patches belonging to the 'real'. In the training, the loss will be calculated following the objective function in Section IV-C.

Besides, we also introduce an image pool to improve the stability of adversarial training by updating the weight in discriminator using samples generated by previous networks. In our experiments, the image pool buffered 4 tuples (including MR image x and its label y), which is denoted by (x_i, y_i) (i is the index in the pool). In each training iteration, we randomly decide whether to exchange samples (x^*, y^*) and (x_i, y_i). If it is necessary, we select the replacer based on a uniform distribution. Although the batch size in our experiment is 1,

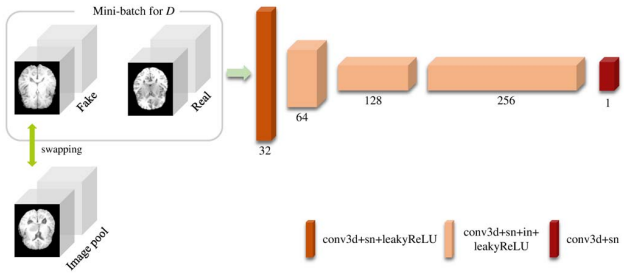

Fig. 4. The discriminator in our network. The input mini-batch can be divided into two groups: Fake and Real, which have the same size. In the Fake group, the images produced by the current generator are randomly replaced by the previous samples from the image pool. Then, the discriminator reduces the size of the input to a quarter by stride 2 convolution in the first two layers and outputs $w \times h \times d$ cubes finally.

this technique is still beneficial. The comparison between 'with pool' and 'without pool' shows in Figure 7.

C. Objective function

The minimax objective function of generative adversarial networks can be formulated as follow:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

In MM-GAN, the objective function used in LS-GAN [37] instead of common GAN loss [10] is applied, since LS-GAN replaces the sigmoid cross entropy loss function to the least squares loss function, which improves the quality of generated images and stabilizes the training process. The objective function is defined as following:

$$\min_D \mathcal{L}_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x,y}[\sum_i^w \sum_j^d \sum_k^h (D(x,y) - b)^2] + \frac{1}{2}\bullet\, {}_x[\sum_i^w \sum_j^d \sum_k^h (D(x, G(x)) - a)^2] \quad (1)$$

$$\min_G \mathcal{L}_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_x[\sum_i^w \sum_j^d \sum_k^h (D(x, G(x)) - c)^2] \quad (2)$$

where a and b label the fake and real data individually, and c is the threshold matrix that G makes D believe the fake data to real. We initialize a and c as $w \times h \times d$ with constant elements 1, and b as a zero matrix with the same size.

V. Experiments

In this section, we evaluate the synthetic image qualitatively and quantitatively with auxiliary segmentation networks and demonstrate the effectiveness in improving the performance of top-performing algorithms.

A. Data Preparation

Experiments are conducted on *BRATS17* datasets [38], and then tentatively extended to the new dataset *LIVER100*. The *BRATS17* dataset has a large number of multimodal MRI scans with high-grade glioblastoma (GBM/HGG) and low-grade glioma(LGG). Each instance consists of 4 modalities MR images (T1 weighted, post-contrast T1-weighted, T2-weighted and FLAIR)[1] that were skull-stripped and co-registered, and the corresponding tumor label annotated by one to four raters manually. It is worth noting that tumors are labeled in 4 parts, necrosis and non-enhancing tumor (label 1), edema (label 2), and active/enhancing tumor (label 4) and remained parts (label 0). The models are required to be trained on its training set containing 210 HGG and 75 LGG, and evaluation results are obtained on its website. The evaluation is carried out with three tasks:

- The whole tumor (necrosis and non-enhancing, enhancing tumor and edema), abbreviated as WT.
- The tumor core (necrosis and non-enhancing and enhancing tumor), abbreviated as TC.
- The enhancing tumor region, abbreviated as ET.

In our experiments in both *BRATS17* and *LIVER100*, the volumes, including MR images and the labels, were cropped to $200 \times 160 \times 150$ from $240 \times 240 \times 155$ pixels to save GPU memory. For *BRATS17*, the intensities of 4 modalities MR images were normalized to have zero mean and unit variance to accelerate the model convergence. Considering that the labels of the tumor can not provide sufficient information to generate MR images, we also label the brain region, other than the entire tumor as label 3. Figure 5 shows the label maps of T1 mode and FLAIR mode and its MR images. Due to the difference of multimodal MR image, each MR image has its respective label map after labeling the brain regions.

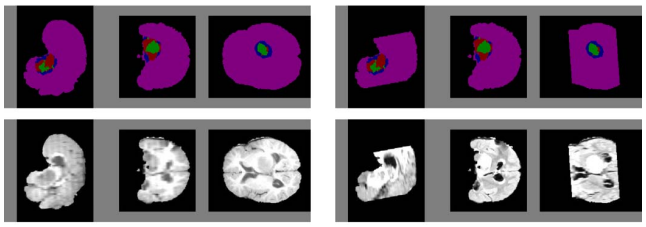

Fig. 5. Examples of label maps in *BRATS17*. Label maps of T1 image and FLAIR image are shown in the top row. The region in red, blue and green denote the necrosis and non-enhancing tumor (label 1), the edema (label 2) and the enhancing tumor (label 4) respectively. The magenta part represents the whole brain except for tumors (label 3). Remained parts of the MR image (label 0) are labeled in black. Although the MR images of mode T1 and mode FLAIR shared the same label map of tumors, the label map of the whole brain is different. The FLAIR image (right of the bottom row) shows relatively small areas of the brain compared to the T1 image (left of the bottom row).

B. Synthesis Evaluation for BRATS17

Here we show the synthetic MR images and assess its quality quantitatively. We trained 4 models for 4 modalities

[1]They are abbreviated as T1, T1ce, T2 and FLAIR.

(T1, T1ce, T2, and FLAIR). For all models, we train for 200 epochs with a minibatch size of 1 [2] and decay the learning rate to 0 at the same step after 100 epochs. These models are optimized by Adam with following hyperparameters: β_1 = 0 and β_2 = 0.9. The default learning rates for generator and discriminator are $4e-4$ and $1e-4$ respectively.

New data can be shared with the public without the worry of privacy leak. Our method can create new data that are not attributed to any patients. Figure 6 shows the representative results of our algorithm. The main structures of the brain, such as the cerebrum, cerebellum, diencephalon, and brainstem, can be identified from the synthetic multi-modal MR images and the fine details, such as sulci and gyri, also recovered partly. The most important thing is that the tumor in our result reflects the information provided by the tumor label map well. Simultaneously, an obvious appearance change between real MR images and synthetic images can be perceived.

Quantitatively, our model achieves very good performance. We introduce four 3D U-Net models of the same architecture as the generator to segment the new data and evaluate the results generated by the real label map, which is a more direct way compared with the common metrics (e.g. IS and FID). Each model is trained for a specific modality (T1, T1ce, T2 and FLAIR) with Adam optimizer (lr = $2e-4$, β_1 = 0.5 and β_2 = 0.999) and the evaluation is conducted with all three tasks (see Figure 7). Besides, the test results of our method without the image pool are also provided in Figure 7. Figure 7 shows the evaluation of the quality of synthetic images by 3D U-Net. As seen in the histograms, the dice score of the Unet and 'with pool' are close, while that of the without pool is relatively poor. From the perspective of FLAIR, it is obvious that the dice score of 'with pool' outperforms that of 'without pool' in the whole tumor, the core tumor and the enhanced tumor.

The same trend can be found in T1ce. While in T1 and T2, and some of the dice scores of 'without pool' are marginally better than that of 'with pool'. All experiments are performed on the train set of BRATS 2017 and the metric is Dice Score [39], which is defined as Eq. 3.

$$s_v = 2\left|x \cdot y\right| / \left(\left|x\right|^2 + \left|y\right|^2\right). \tag{3}$$

where x denotes the binary map of the segmentation result of synthetic images and y is the one-hot map of ground truth.

The following conclusion can be drawn:

1) The synthetic images are acceptable and reliable, and the dice scores of the whole tumor in four modalities are approximate to 0.8. Due to class imbalance and insufficient data, the relatively small area of the enhancing tumor is failed to be restored well. Its scores of T1, T2 and FLAIR MR images are below 0.4.
2) Remarkably, the fake images in T1ce mode hit the best overall performance in three tasks. This further shows the efficacy of our model.
3) Using an image pool is beneficial from this mechanism, except T2 (which only drops 0.01).

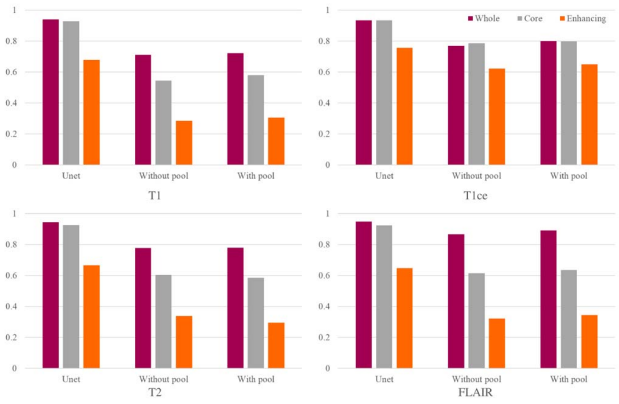

Fig. 7. Evaluation of the quality of synthetic images by 3D U-Net. The legend located on the top right and the vertical axis represents the dice score.

C. Data Augmentation in Segmentation

Through perturbing the existing label maps randomly, we have acquired new label maps, which are then converted into multi-modal MR images by *MM-GANs*. Such a process will provide abundant data for training segmentation networks, i.e. the multi-modal MR images and its label map. More concretely, we transform the tumors in label maps by deformation, flip, rotate, and scale, instead of the whole label map. And all alterations are performed on T1 mode's label maps in our experiments since the T1 images record the whole region of brains perfectly (Figure 5).

We further build a new dataset based on the training set. The new set containing 4 modalities is used to train a variant of 3D U-Net [3] and Triple-Cascaded-Net [40]. The latter is the 2nd place of BRATS 2017 Challenge, which uses the cascaded three binary segmentation networks by the hierarchy of subregion of the tumor. The first network segments the whole tumor (including labels 1, 2 and 4) from the entire image and crops its bounding box area as the next network input. The tumor core (including labels 1 and 4) is segmented by the second network based on this input. And its bound box area also provides to the next network to segment the subregion of enhancing tumor (including the label 4).

Table I shows the results of the networks trained on different datasets constructed from BraTS 2017. All results are obtained by uploading the test results of the BraTS 2017 validation set [4] to its official website. There are five settings of the training set respectively as: (1) 100% real images, (2) 100% fake images, (3) 100% fake images and the mix of 100% real images, (4) 300% fake images and the mix of 100% real images, and (5) 500% fake images and the mix of 100% real images.

As we expect, the pure fake data set has the worst performance due to the conflict between the four modalities and its

[2]It's the largest batch size in Nvidia GTX 1080Ti GPUs

[3]The codes source from https://github.com/ellisdg/3DUnetCNN.

[4]The test set is not accessible except the competition participants.

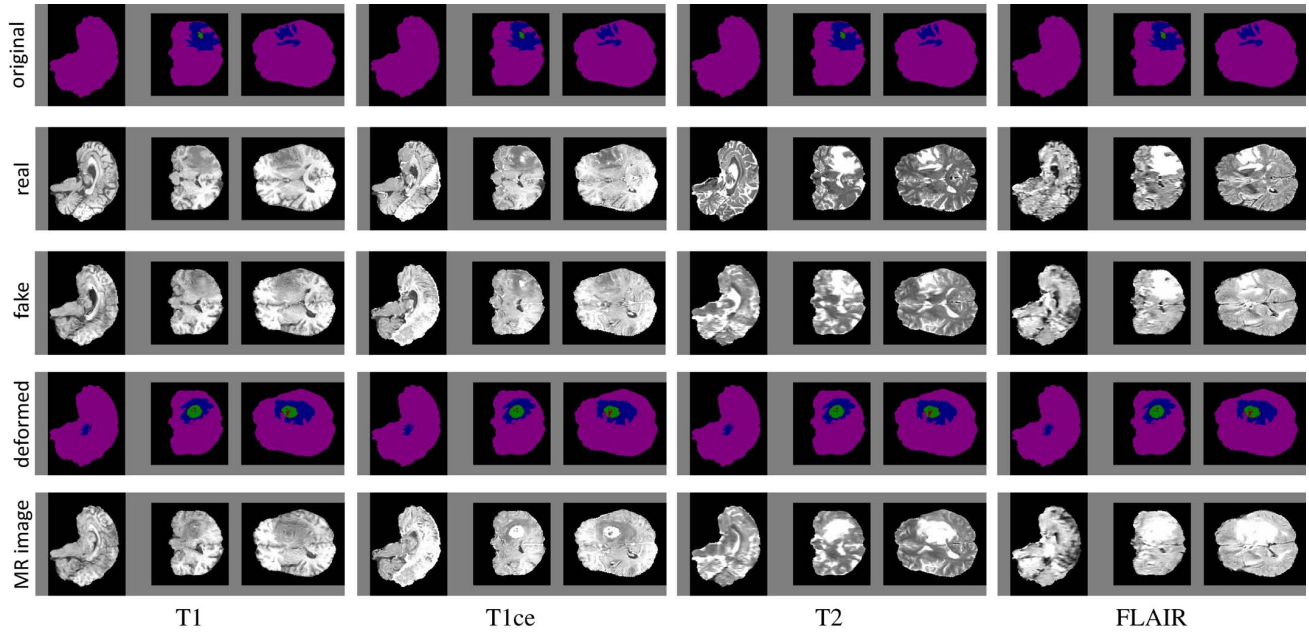

Fig. 6. Example of synthetic data. The first three rows are original label maps of the brain and its real MR images and synthetic images. Remaining two rows are label maps converted from the first row and its synthetic image. From left to right, the modalities of columns are T1, T1ce, T2 and FLAIR. All images are represented by the center slices of the sagittal, coronal and axial planes.

poor details. As is shown on Table I and Figure 8, the top half of the table are the results of the model based on 3D U-Net.

From the perspective of the whole tumor(WT), the result of the networks trained on 500% fake images and the mix of 100% real images has increased 1.9 percent than 100% real images. While that of the enhanced tumor(ET) has slightly increased 0.4 percent. The result of the tumor core(TC) shows that 300% fake images and the mix of 100% real images of perform the best, which has increased 2.1 percent than 100% real images. Compared with the top half, the bottom half of Table I and Figure 9 are the results of Triple-Cascaded-Net [40], which have increased weekly after adding the fake data. The result of the whole tumor shows that the networks trained on 500% fake images and the mix of 100% real images have increased 0.01 percent than 100% real images. The result of the enhanced tumor has the same trend, that the networks trained on 500% fake images and the mix of 100% real images has increased 0.4 percent than 100% real images. However, in the tumor core, the addition of the fake image data doesn't exceed the original results of 100% real images.

The attractive results are shown on 3D U-Net: the dice scores of the whole tumor and tumor core are increased by at most 0.017 and 0.016 respectively. On the contrary, the mixed training set exhibits comparable performance on Wang et.al. [40] with pure real data set. It is mainly because Triple-Cascaded-Net is more focused on local information, while 3D U-Net combines global and local information.

Furthermore, to demonstrate the effect of the fake data, we also use 10% (i.e. 29) real data to fine-tune the model trained on fake data set and plot its result in Figure 10. As can

Data Component	WT	TC	ET
100% R	0.8722±0.14	0.7525±0.24	0.6638±0.33
100% F	0.6539±0.25	0.5170±0.28	0.4117±0.34
100% F + 100% R	0.8720±0.14	0.7621±0.21	0.6632±0.33
300% F + 100% R	0.8875±0.09	**0.7685±0.20**	0.6618±0.33
500% F + 100% R	**0.8891±0.10**	0.7674±0.20	**0.6662±0.32**
100% R	0.8989±0.07	**0.8356±0.16**	0.7349±0.29
100% F	0.5227±0.30	0.3763±0.30	0.3380±0.31
100% F + 100% R	0.8981±0.07	0.8249±0.15	0.7321±0.28
300% F + 100% R	0.8963±0.08	0.8298±0.16	0.7346±0.28
500% F + 100% R	**0.8990±0.08**	0.8305±0.15	**0.7379±0.27**

TABLE I
PERFORMANCE OF THE SEGMENTATION NETWORKS TRAINED BY THE DIFFERENT DATA COMPONENTS GAINED FROM *BRATS17*. THE FIRST PART ARE THE RESULTS OF MODEL BASED ON 3D U-NET, AND THE SECOND ARE THOSE OF TRIPLE-CASCADED-NET. HERE, THE F IS THE ABBREVIATION OF FAKE DATA, AND R IS THE ABBREVIATION OF REAL DATA. AN OBVIOUS INCREASE IS PRESENTED IN 3D U-NET.

be observed, even fine-tuning the fake data model with only 10% real data, the result also approximates the model trained by the complete real data and far better than the original model. Therefore, we can use anonymous fake data to pre-train the network and fine-tune it with a small amount of real data without worrying about performance degradation, which protects the patient information and greatly cuts down the cost in data collection.

D. Synthesis Evaluation for LIVER100

Due to the limited data size, the *LIVER100* is more suitable and challenging to implement data augmentation. In the ex-

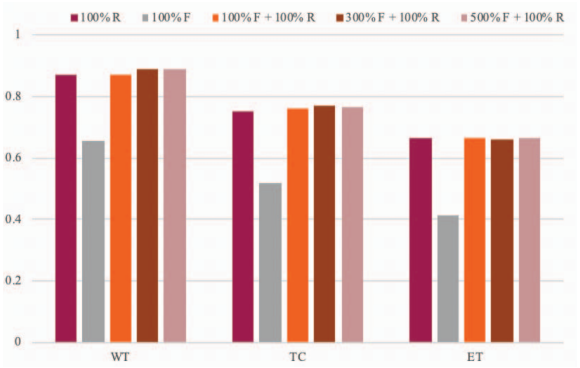

Fig. 8. Performance of 3D U-Net trained by different data components from *BRATS17*.

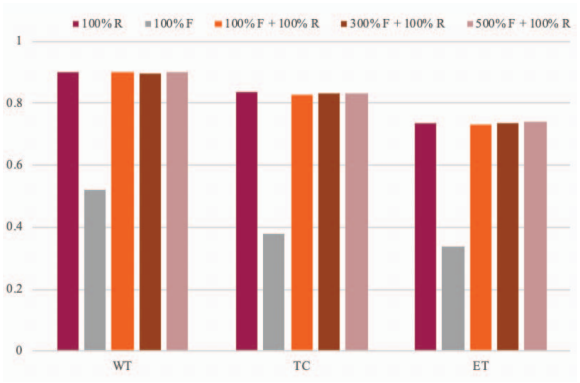

Fig. 9. Performance of Triple-Cascaded-Net trained by different data components from *BRATS17*.

periment, we separate the dataset randomly into the training set and test set, according to the proportion 4:1 respectively.

Similar to *BRATS17*, we have trained the segmentation model under five settings of the training set respectively as: (1) 100% real images, (2) 100% fake images, (3) 100% fake images and the mix of 100% real images, (4) 300% fake images and the mix of 100% real images, and (5) 500% fake images and the mix of 100% real images.

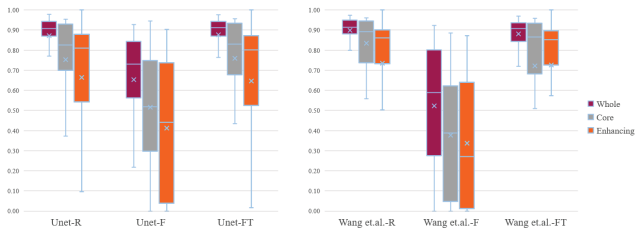

Fig. 10. Comparison of the real data model and the fine-tuning model. The left figure plotted the performance of the 3D U-Net trained by different means. 'Unet-R' and 'Unet-F' are trained by real data and fake data respectively. 'Unet-FT' is the result fine-tuned on 'Unet-F' by 10% real data. The right figure shows the results of the above training methods on the Wang et.al.'s model.

The model are trained based on 3D U-Net for 100 epochs with a minibatch size of 1, and we decay the learning rate to 0 after 50 epochs. The model is also optimized by Adam with β_1 = 0 and β_2 = 0.9. The default learning rates for generator and discriminator are set to $4e-5$ and $1e-5$ respectively. Table II illustrates the performance of our model on the *LIVER100*.

Table II and Figure 11 are the performance reslut of the segmentation networks trained by the different data components gained from *LIVER100*. The tumor dice score of the 3D U-Net with 300% fake images and 100% real images were increased by 0.7%, 65.7%, 1.6%, 0.5%, compared with 3D U-Net with 100% real data, 100% fake images, 100% fake images and 100% real images, and 500% fake images and 100% real images. We can see from Table II that compared with the dice score of 100% real images, the performance of the networks trained on 500% fake images and the mix of 100% real images has increased 0.7 percent. The same as in *BRATS17*, the pure fake data set has the worst performance. There is, by comparison, a mere 0.5% growth in the performance occurs under setting (4) (mixed training data of 300% fake images and 100% real images). Insufficient computing resources restrict the fine-tuning process, so we believe the model can be optimized further on *LIVER100* afterward.

Data Component	Dice Score
100% R	0.6856±0.18
100% F	0.4166±0.30
100% F + 100% R	0.6795±0.19
300% F + 100% R	**0.6903±0.20**
500% F + 100% R	0.6880±0.18

TABLE II
PERFORMANCE OF THE SEGMENTATION NETWORKS TRAINED BY THE DIFFERENT DATA COMPONENTS GAINED FROM *LIVER100*. THE SEGMENTATION MODEL IS BASED ON 3D U-NET. COMPARED WITH (1), THE DICE SCORE RISES A SLIGHT 0.5 PERCENT IN (4). HERE, THE F IS THE ABBREVIATION OF FAKE DATA, AND R IS THE ABBREVIATION OF REAL DATA.

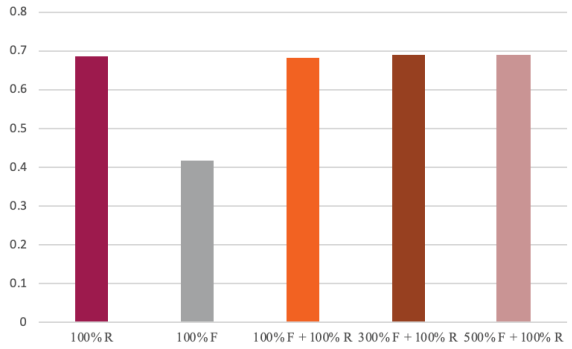

Fig. 11. Performance of 3D U-Net trained by different data components from *LIVER100*.

VI. CONCLUSION

In this paper, we propose a generative model *MM-GAN* to synthesize 3D MR images of brain tumors and liver lesions

from a deformed label map using the GANs. It can produce abundant data with labels to promote segmentation networks. Besides, *MM-GAN* also can hide the patients' information effectively and facilitate data sharing. Compared to [12], the synthesis result can be controlled by modifying the input label map, but a fine label map is not necessary and the pipeline is much shorter. The augmentation and segmentation model are transferred to the *LIVER100* dataset, and the results show an initial growth in the performance of liver lesion segmentation.

REFERENCES

[1] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. Li, and L. Feifei, "Thoracic disease identification and localization with limited supervision," *CVPR*, 2018.

[2] J. Yang, X. Sun, J. Liang, and P. L. Rosin, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *CVPR*, 2018.

[3] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *CVPR*, 2018.

[4] B. Teixeira, V. Singh, T. Chen, K. Ma, B. Tamersoy, Y. Wu, E. Balashova, and D. Comaniciu, "Generating synthetic x-ray images of a person from the surface geometry," in *CVPR*, 2018.

[5] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *CVPR*, 2018.

[6] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *JMI*, 2018.

[7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *CVPR*, 2017.

[8] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *TMI*, 2010.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[11] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv:1711.04340*, 2017.

[12] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging, Workshop*, 2018.

[13] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *CoRR*, vol. abs/1701.07875, 2017.

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[15] H. Zhang, T. Xu, and H. Li, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *ICCV*, 2017.

[16] A. Antoniou, A. J. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *ICLR*, 2018.

[17] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. A. Sutton, "VEEGAN: reducing mode collapse in gans using implicit variational learning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 3308–3318.

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *CVPR*, 2018.

[19] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *NIPS*, 2016.

[20] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv:1802.05957*, 2018.

[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017.

[22] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *NIPS*, 2016.

[23] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," *arXiv:1609.07093*, 2016.

[24] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE TMI*, 2018.

[25] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *MICCAI*, 2017.

[26] J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan, "Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *MICCAI*, 2018, pp. 777–785.

[27] P. Zhang, F. Wang, W. Xu, and Y. Li, "Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in k-space," in *MICCAI*, 2018.

[28] Y. Hu, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, T. Vercauteren, J. A. Noble, and D. C. Barratt, "Adversarial deformation regularization for training image registration neural networks," *arXiv:1805.10665*, 2018.

[29] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, "Adversarial similarity network for evaluating image alignment in deep learning based registration," in *MICCAI*, 2018.

[30] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho, "Towards adversarial retinal image synthesis," *arXiv:1701.08974*, 2017.

[31] T. C. Mok and A. C. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," *arXiv:1805.11291*, 2018.

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *MICCAI*, 2016.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[34] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, 2016.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv:1805.08318*, 2018.

[36] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," *CVPR*, 2017.

[37] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *ICCV*, 2017.

[38] 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 TMI*, 2015.

[39] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, 1945.

[40] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," *MICCAI*, 2017.

[41] B. Wang, S. Qiu, and H. He, "Dual encoding u-net for retinal vessel segmentation," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 11764, pp. 84–92.

[42] X. Guo and Y. Yuan, "Triple anet: Adaptive abnormal-aware attention network for WCE image classification," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 11764. Springer, 2019, pp. 293–301.