Harmonizing Pathological and Normal Pixels for Pseudo-Healthy Synthesis

Yunlong Zhang, Xin Lin, Yihong Zhuang, Liyan Sun, Yue Huang, *Member, IEEE*, Xinghao Ding, Guisheng Wang, Lin Yang, and Yizhou Yu, *Fellow, IEEE*

***Abstract*— Synthesizing a subject-specific pathology-free image from a pathological image is valuable for algorithm development and clinical practice. In recent years, several approaches based on the Generative Adversarial Network (GAN) have achieved promising results in pseudo-healthy synthesis. However, the discriminator (i.e., a classifier) in the GAN cannot accurately identify lesions and further hampers from generating admirable pseudo-healthy images. To address this problem, we present a new type of discriminator, the segmentor, to accurately locate the lesions and improve the visual quality of pseudo-healthy images. Then, we apply the generated images into medical image enhancement and utilize the enhanced results to cope with the low contrast problem existing in medical image segmentation. Furthermore, a reliable metric is proposed by utilizing two attributes of label noise to measure the health of synthetic images. Comprehensive experiments on the T2 modality of BraTS demonstrate that the proposed method substantially outperforms the state-of-the-art methods. The method achieves better performance than the existing methods with only 30% of the training data. The effectiveness of the proposed method is also demonstrated on the LiTS and the T1 modality of BraTS. The code and the pre-trained model of this study are publicly available at https://github.com/Au3C2/Generator-Versus-Segmentor.**

***Index Terms*— Medical image synthesis, low-contrast medical image segmentation, adversarial training, image enhancement, label noise.**

Manuscript received 20 November 2021; revised 16 January 2022 and 26 February 2022; accepted 24 March 2022. Date of publication 1 April 2022; date of current version 31 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC0118101; in part by the National Natural Science Foundation of China under Grant 82172033, Grant 52105126, Grant U19B2031, and Grant 61971369; in part by the China Postdoctoral Science Foundation under Grant 2021M702726; and in part by the Science and Technology Key Project of Fujian Province under Grant 2019HZ020009. *(Corresponding author: Yue Huang.)*

Yunlong Zhang was with the School of Informatics, Xiamen University, Xiamen 361005, China. He is now with the School of Engineering, Westlake University, Hangzhou 310012, China (e-mail: 23320181154356@stu.xmu.edu.cn).

Xin Lin, Yihong Zhuang, Yue Huang, and Xinghao Ding are with the School of Informatics, Xiamen University, Xiamen 361005, China, and also with the Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: joeylin@stu.xmu.edu.cn; zhuangyihong@stu.xmu.edu.cn; huangyue05@gmail.com; dxh@xmu.edu.cn).

Liyan Sun is with the School of Electronic Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: sunly@stu.xmu.edu.cn).

Guisheng Wang is with the Department of Radiology, Third Medical Centre, Chinese PLA General Hospital, Beijing 100853, China (e-mail: wanggs1996@tom.com).

Lin Yang is with the School of Engineering, Westlake University, Hangzhou 310012, China (e-mail: yanglin@westlake.edu.cn).

Yizhou Yu is with the Deepwise AI Laboratory, Beijing 100125, China, and also with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: yizhouy@acm.org).

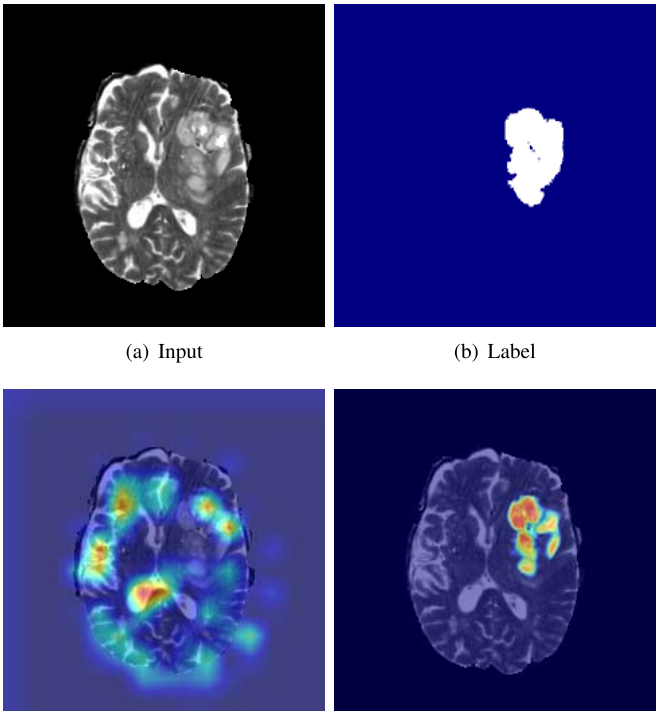
Digital Object Identifier 10.1109/TMI.2022.3164095

I. Introduction

PSEUDO-HEALTHY synthesis is defined as synthesizing a subject-specific pathology-free image from a pathological image [1], [2]. Generating such images has been demonstrated to be valuable for a variety of tasks in medical image analysis [2], such as segmentation [1], [3]–[7], detection [8], and assisting doctors with diagnosis by comparing the pathological and pseudo-healthy images [5], [8], [9]. *By definition, a perfect pseudo-healthy image should maintain healthiness (i.e., synthesizing healthy-like appearances) and subject identity (i.e., belonging to the same subject as the input). Both attributes are essential and indispensable. The former attribute is self-explanatory, while the latter one is also considerable since generating another healthy image is meaningless.*

Pseudo-healthy image synthesis is an ill-posed inverse procedure since there exists a multitude of healthy-looking solutions for a pathological input. To tackle this inverse problem, several GAN-based methods [2], [5], [9] have been presented. The basic architecture in these methods includes a generator and a discriminator. The generator is an encoder-decoder architecture trained to translate pathological images into corresponding healthy-looking ones, whereas the discriminator competes against the generator and aims to differentiate the synthetic and healthy images by a two-way classifier. However, choosing the classifier as the discriminator has the shortcoming that lesions cannot be accurately located (see Fig. 1(c)) due to the following reasons: (1) The visual explanations of the classifier involve healthy regions, which will further lead to falsely erasing the subject identity. (2) The highlighted explanations of the classifier cannot cover the entire tumor region, which easily causes the pathology remains in the synthetic images.

To address the problem of inaccurate localization, we choose to use a segmentor as the discriminator. The segmentor identifies pathological regions more accurately than the classifier (e.g., the 'tumor' explanation accurately highlights the tumor regions in Fig. 1(d)). Therefore, both the subject



(a) Input (b) Label

(c) Classifier (d) Segmentor

Fig. 1. 'Visual explanations' of the classifier and the segmentor. (a-d) represent the input, tumor annotation, class activation maps of the classifier generated by Grad-CAM [10], and class activation maps of the segmentor generated by and Seg-Grad-CAM [11], respectively.

identity and healthiness can be simultaneously guaranteed by keeping the healthy pixels and transforming the pathological ones, respectively. The former is implemented by a visual consistency loss, and the latter is achieved by the adversarial training between the generator and the segmentor.

During the process of alternatively training the generator and the segmentor, the pathological pixels are gradually transformed into healthy-looking pixels. As a result, these healthy-looking pixels (i.e, should be labeled as 'no tumor') are falsely labeled as 'tumor' when training the segmentor, which results in the poor generalization of the segmentor and further hampers the entire training process. To resist false labels as much as possible, we propose a difference-aware loss to improve the generalization ability of the segmentor by muting those deceptively healthy-looking pixels.

Moreover, we also develop one downstream task of pseudo-healthy synthesis that increases the contrast between lesions and normal tissue. It is well-known that the low contrast, especially for lesions, is a nasty nature of medical images that may result in miss detection of low-contrast anatomies in medical image segmentation. Since the proposed method can effectively disentangle a disease image into its two components, the healthy and the pathological parts that only contain healthy tissue information and pathological information, respectively. We design an image enhancement technique that highlights lesions by simply adding the extracted pathological signals into the inputs. Then, this technique is applied to alleviate the low contrast problem for medical image segmentation.

Lacking a good quantitative metric for measuring the healthiness is one of the major barriers in pseudo-healthy synthesis. Recently, Xia *et al.* [2] proposed a metric using a pre-trained segmentor. However, this metric has *poor stability (i.e., fluctuating drastically at multiple trials) and vulnerability (i.e., predicting falsely when synthetic pathology slightly deviates from the true pathology but is apparently abnormal)*. Besides, the subjective assessment was viewed as the gold standard to ultimately determine the healthiness [2]. However, it is time-consuming, costly, and subject to inter- and intra-observer variability, and hence also deviates from the reproducibility. Inspired by two attributes of the label noise that fitting incorrect labels requires more time [12] and the higher learning rate hampers the memorization of false labels [13], we propose a reliable metric for measuring the healthiness by estimating the convergence speed.

In order to evaluate the effectiveness of the proposed approach, we conduct extensive experiments on the T2 modality of BraTS dataset [14]–[16]. Furthermore, we also present a part of results on the T1 modality of BraTS and the LiTS datasets [17] to demonstrate that the adaptive capability of our method to other modalities and organs.

This work is a significant extension of our prior conference paper [18]. The main contributions are summarized as follows, (1) This paper further clearly and intuitively unravels the motivation of the Generative versus Segmentor (GVS) by using the class activation mapping technique. (2) This paper proposes a healthiness metric by addressing the instability of the $\mathbb{S}_{dice}$ in the conference version. Then, the reliability of the improved metric is validated by multiple experiments. Besides, this paper introduces the subjective assessments to further evaluate the effectiveness of the proposed method. (3) This paper develops the application of pseudo-healthy synthesis for image enhancement. The enhanced images not only contribute to identifying lesions in visual but also improve lesion segmentation performance.

II. Related Works

A. Pseudo-Healthy Synthesis

Recently, pseudo-healthy synthesis has attracted considerable attention in the medical image analysis community because of its potential for downstream tasks [2]–[6], [8], [9], [19]. The related work is divided into two main categories, pathology-deficiency (i.e., only providing healthy images in the training phase) [19]–[21] and pathology-sufficiency based methods (i.e., possessing plenty of pathological and healthy images in the training phase) [2], [5], [9].

The pathology-deficiency based methods [19], [22]–[28] were always closely associated with unsupervised anomaly detection/segmentation [29], which aimed to learn normative distributions by learning to compress and recover healthy anatomies in the training phase. In the subsequent testing stage, pathological images were first compressed to the latent space. Then, pseudo-healthy images were reconstructed from latent representations based on the assumption that the obtained latent representations were close to the latent representations of pseudo-healthy images. However, the

assumptions of these methods were too idealistic [19], [21]. Actually, the healthiness and the subject identity were not guaranteed due to the difficulty in finding the optimal latent representations, corresponding to pseudo-healthy images when compressing the pathological images into latent space.

The pathology-sufficiency based methods [2], [5], [9] tackled pseudo-healthy synthesis from the viewpoint of image translation. In the training phase, these methods introduced pathological images alongside corresponding image-level [9] or pixel-level [2], [5] pathological annotations to learn an image translation process of mapping pathological images to pseudo-healthy images. Baumgartner *et al.* [9] proposed a basic GAN-based scheme, that was composed of a generator and a discriminator. The generator was trained to synthesize healthy-looking appearances and keep the subject identity at the same time, whereas the discriminator aimed to differentiate the synthetic images from the unpaired healthy images. This method only used the image-level annotations and was not able to accurately translate pathological pixels and keep the healthy ones. To alleviate these issues, PHS-GAN [2] and ANT-GAN [5] introduced pixel-level annotations. Both methods were variants of Cycle-GAN [30]. Specifically, the PHS-GAN considered the one-to-many problem and disentangled the information of pathology from what seems to be healthy, and pixel-level labels were used to extract the location and shape of pathology. In the process of applying Cycle-GAN to pseudo-healthy synthesis, the ANT-GAN proposed the shortcut to simplify the optimization and the masked L2 loss to better preserve the normal regions.

Our experimental setting is identical to the work by Sun *et al.* [5] and Xia *et al.* [2]. However, our motivation is significantly different from theirs. They tried to translate the pathological images into healthy-looking appearances. In comparison, the proposed method further explicitly utilizes the information inside the appearance differences between healthy and pathological regions and tries to generate pathology-free images by making up such differences until achieving harmony between them.

B. Adversarial Training

The idea of adopting the segmentor as the discriminator is inspired by the extensive applications of adversarial training [31]. The discriminator of the original GAN [31] was used to differentiate true or counterfeit images. In domain adaptation, DANN [32] adopted a discriminator to distinguish the data sampled from the source or the target domain. In the research on the adversarial attack, the discriminator is a classifier to sort an example into a corresponding class [33], [34]. In image translation tasks, the discriminator is also a classifier to detect the high-level structured differences between the translated and real images [30], [35]. Recently, Naveen *et al.* [36] applied adversarial training into self-supervised learning and adopted a classifier as the discriminator to predict pretext labels. Compared with the related work using diverse classifiers as discriminators, this paper further develops the paradigm of adversarial training and extends the discriminator to pixel-level dense prediction tasks.

C. Medical Image Enhancement

It is well-known that the low contrast is an intrinsic nature of medical images, which will hinder the clinical decision-making and the downstream analysis tasks (e.g., segmentation, detection). To alleviate this issue, image enhancement has been widely studied in the field of medical imaging [37]. Singh *et al.* [38] and Wang *et al.* [39] improved the histogram equalization, the most basic enhancement technique, to improve the visual quality of low radiance retinal images and multiple types of medical images. Frosio *et al.* [40] enhanced the digital cephalic radiography with mixture models and local gamma correction. Hamghalam *et al.* [41] proposed an image-to-image translation technique to generate synthetic high tissue contrast (HTC) images and used the enhanced images to improve the segmentation performance. However, most of the existing studies did not consider the context. Hence, they can easily fail when encountering complex contexts. For example, the cerebrospinal fluid and the tumor both exhibit high signal intensity on T2-weighted images of BraTS dataset. The existing methods cannot identify them and increase the contrast between them, which is harmful to identify the lesion after enhancement. Compared with the existing methods, the proposed enhancement method considers the context of images and can adaptively enhance the contrast between lesions and normal tissues.

III. Methods

The proposed Generative versus Segmentor (GVS) for pathology-sufficiency pseudo-healthy synthesis with pixel-level labeling is introduced in Sec. III-A and III-B. Then, in Sec. III-C, we apply synthetic images to medical image enhancement. Furthermore, the enhanced images are used to help low-contrast medical image segmentation.

Assuming that a set of pathological images $\{x_p\}$ with their pixel-level lesion annotations $\{y_t\}$ are given. Our goal is to train a generator $\mathbf{G}$ that can translate the pathological image x_p into a corresponding synthetic image x_s with superior healthiness and subject identity.

A. Basic GVS Flowchart

The training workflow of the proposed GVS is shown in Fig. 2. The generator gradually synthesizes the healthy-looking images by iteratively alternating Steps A and B. The specific steps are described as follows.

Step A: As shown in Step A of Fig. 2, we fix the generator $\mathbf{G}$ and update the segmentor $\mathbf{S}$ to detect the lesions in the synthetic images. To this end, the segmentation loss used to train the segmentor is defined as:

$$\mathcal{L}_{s1} = \mathcal{L}_{ce}(\mathbf{S}(x_s)), y_t), \tag{1}$$

where L_{ce} denotes the cross-entropy loss. Otherwise, the y_t are binary labels, where 0 and 1 represent the normal and the pathological regions, respectively.

Step B: In this step, we fix the segmentor $\mathbf{S}$ and update the generator $\mathbf{G}$, aiming to remove the lesions and preserve the subject identity of the pathological images.

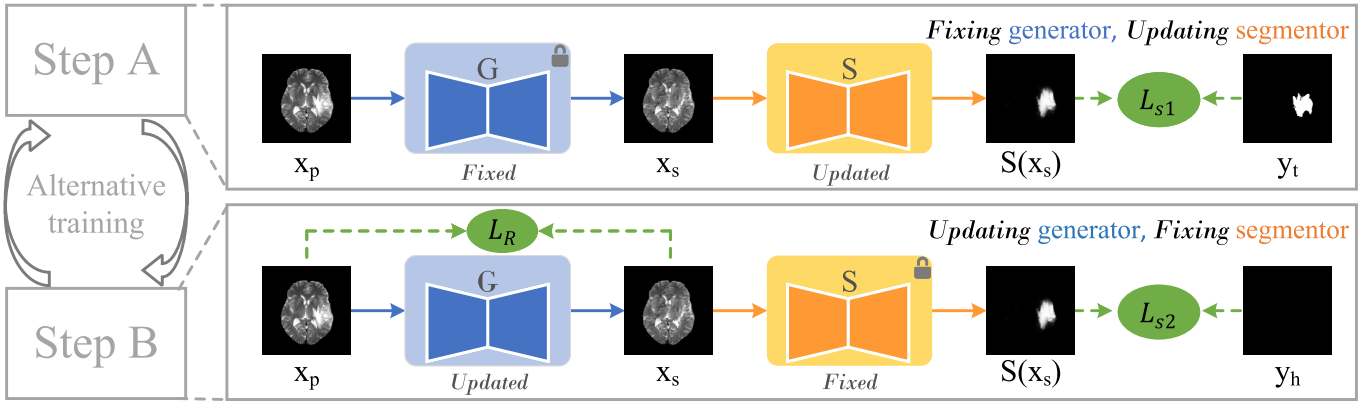


Fig. 2. Training workflow. The model is optimized by iteratively alternating Step A and Step B (left chart). In Step A (right top), we fix the generator **G** and update the segmentor **S** with L_{s1}. In Step B (right bottom), we fix the segmentor **S** and update the generator **G** with $L_{s2} + \lambda L_R$.

Firstly, it is expected that the generator **G** can synthesize healthy-looking images. To achieve this, the generator is trained to deceive the segmentor by compensating the appearance difference between pathological and healthy regions. Specifically, another segmentation loss is adopted as follows:

$$\mathcal{L}_{s2} = \mathcal{L}_{ce}(\mathbf{S}(\mathbf{G}(x_p))), y_h), \tag{2}$$

where y_h denotes a zero matrix with the same size as y_t.

Secondly, the synthetic images should keep the same subject identity with inputs. Hence, the residual loss proposed in the existing pseudo-healthy synthesis method [5] is used:

$$\mathcal{L}_R = \mathcal{L}_{mse}(x_p, \mathbf{G}(x_p)), \tag{3}$$

where $\mathcal{L}_{mse}$ denotes the pixel-wise $\mathcal{L}_2$ loss. The total training loss of **G** is:

$$\mathcal{L}_G = \mathcal{L}_{s2} + \lambda \mathcal{L}_R, \tag{4}$$

where λ denotes a hyperparameter that trades off the healthiness against subject identity, and $\lambda > 0$.

During the iterative training, the segmentor **S** and the generator **G** compete with each other. The segmentor **S** tries to detect the differences between normal and pathological regions, whereas the generator **G** tries to compose them. Eventually, the generator **G** bridges the gap between the pathological and normal regions and synthesizes healthy-looking images.

B. Difference-Aware Loss

In this section, the generalization ability of the segmentor is further considered. During the training process, the pathological regions are gradually transformed into normal regions. As shown in the yellow box of Fig. 3(b), a major part of the pathological region has been well transformed, so these pixels should be labeled as 'no tumor'. However, the basic GVS still considers these pixels as 'tumor' when training the segmentor, which misguides the segmentor and strongly harms its generalization ability [12]. The green box in Fig. 3(b) shows that the predictions of the segmentor severely deviate from the labels, suggesting the poor generalization ability of the segmentor. To overcome this challenge, we adopt an easy yet effective strategy that is muting the well-transformed pixels. It is discovered that the difference maps between inputs and synthetic outputs can reflect how well the pixels are transformed. That is, the substantial differences denote the good transformation, while the minor differences denote the poor transformation (rf. Fig. 3(a)). Hence, the difference maps are utilized as indicators to measure the transformation degree and a difference-aware loss is proposed for alleviating the overfitting when training the segmentor, which is defined as:

$$\mathcal{L}_{wce} = \frac{1}{N}\sum_{i=1}^{N} w(i) y_t(i) log(\mathbf{S}(\mathbf{G}(x_p))(i)), \tag{5}$$

where N denotes the number of pixels. The weights w associated with difference maps are defined as:

$$w = \begin{cases} 0.1, & 1 - m < 0.1, \\ 1 - m, & \text{Otherwise}, \end{cases} \tag{6}$$

where $m = \text{Normalization}(x_p - \mathbf{G}(x_p))$ denotes the normalized difference map. In this work, $w[w < 0.1] = 0.1$ because the minimum value does not represent perfect transformation, and it is necessary to keep a subtle penalty. The complete GVS is proposed by upgrading the segmentation loss $\mathcal{L}_{s1}$ in Equation 2 to the difference-aware loss $\mathcal{L}_{wce}$ in Equation 5.

C. Lesion Contrast Enhancement and Downstream Segmentation

After the training process, the proposed GVS can effectively disentangle a disease input into healthy (i.e., only containing healthy tissue information) and pathological (i.e., only containing pathological information) parts. Our enhancement is implemented by adding the extracted pathological part into the input, which is formulated as:

$$x_{en} = x_p + \alpha * (x_s - x_p), \tag{7}$$

where α denotes the degree of enhancement and $x_s - x_p$ represents the pathological residue between the disease and pseudo-healthy images. The enhanced images x_{en} increase the intensity of lesions while keeping the normal tissues well when the $x_s - x_p$ only contains the pathological information. The

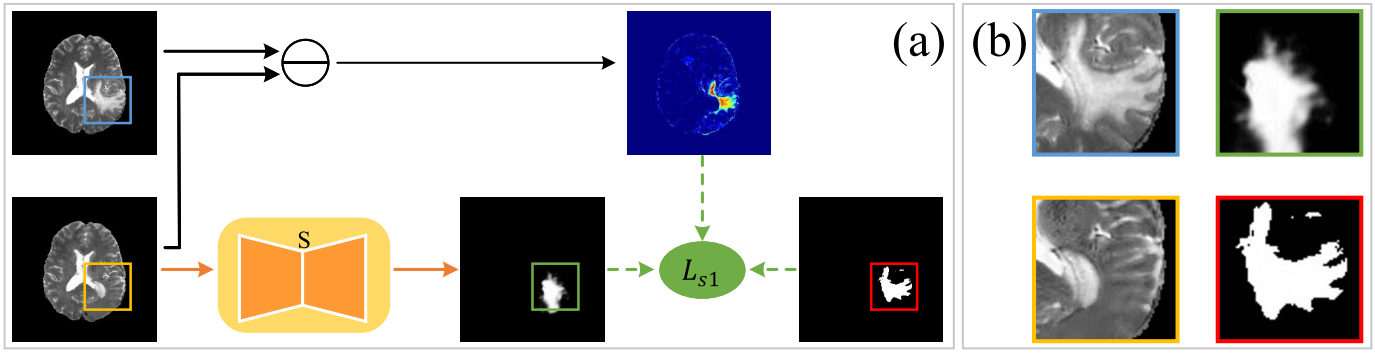
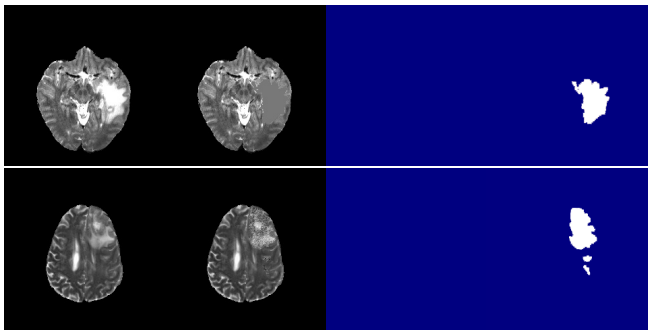


Fig. 3. (a) Framework of the pixel-level weighted cross-entropy loss. (b) The blue, yellow, green, and red boxes denote the pathological image, synthetic image, prediction, and lesion annotation, respectively.



(a) Inputs (b) Counterfeit images (c) Predictions (d) Labels

Fig. 4. Two types of counterfeit images fooling the pre-trained segmentor in literature [2]. Four columns from left to right represent inputs, counterfeit images, the predictions of counterfeit images, and lesion annotations, respectively. The counterfeit image in the first row is generated by filling the pathological regions with the average value of normal tissues. The second row adds the Gaussian noise with zero mean and 0.2 covariance in the pathological regions. The pre-trained segmentor cannot detect the abnormalities existing in both counterfeit images.

proposed GVS can achieve this well and thus can effectively enhance the contrast between the lesions and normal tissues.

Recently, Hamghalam *et al.* [41] revealed that increasing the contrast between tissues can effectively improve the generalization ability of the segmentation task on the BraTS dataset. Inspired by this, we apply the enhanced images to improve lesion segmentation performance. Specifically, the downstream segmentor S_D[1] is trained on the enhanced images. In the subsequent test phase, test samples are also enhanced before being fed into the segmentor.

IV. Measuring Healthiness

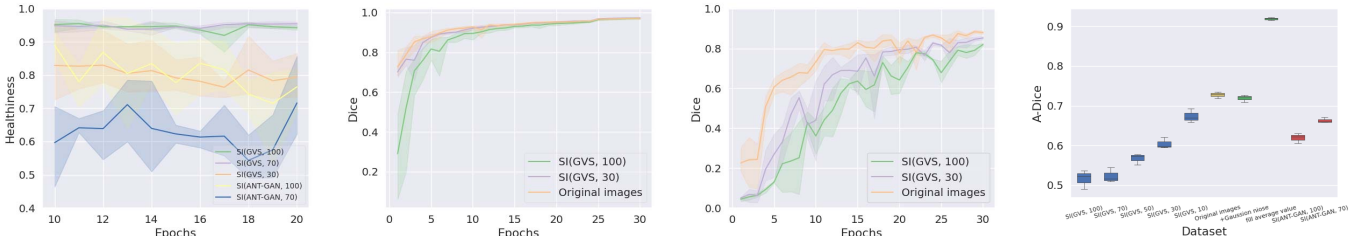
In this section, we analyze the "Healthiness" metric proposed by Xia *et al.* [2] in the subsequent second paragraph. In the third paragraph, we introduce the motivation of the proposed healthiness metric. In the fourth paragraph, we propose a new metric A-Dice to measure the healthiness based on $\mathbb{S}_{dice}$. In the last two paragraphs, the reliability of the proposed metric is validated by analysis and experiments.

In the evaluation process of "Healthiness", a segmentor is first pre-trained to estimate pathology in images. Then the pre-trained segmentor is used to assess pathology within synthetic images by checking how large the estimated pathological areas are. Further implementation details can be found in Sec. 4.4 of literature [2]. This process seems reasonable but cannot guarantee accurate measurement due to the following two shortcomings. First, the pre-trained segmentor is vulnerable to artifacts that are far away from both pathological and healthy appearances. For example, Fig. 4(b) shows two counterfeit images with different artifacts that are clearly abnormal. However, the pre-trained segmentor cannot recognize the abnormalities, resulting in the false high "Healthiness". Second, this metric drastically fluctuates at multiple trials. To verify this, we repeatedly pre-train three segmentors and then use them to calculate the "Healthiness" on the same synthetic images. As shown in Fig. 5(a), the results fluctuate drastically at various epochs and runtimes, especially when the "Healthiness" is smaller (e.g., SI(GVS, 30),[2] SI(ANT-GAN, 100), and SI(ANT-GAN, 70)). Another phenomenon is that the performance of the SI(GVS, 70) will surpass the SI(GVS, 100) at some epochs (e.g., 17^{th} epoch), which is unreasonable since the model trained on more data is superior to that trained on fewer data with a high probability. A similar phenomenon also appears in another pseudo-healthy synthesis method, ANT-GAN.

To measure the healthiness accurately, we propose a new metric, A-Dice, by estimating the convergence speed of fitting an evaluation segmentor S_E to synthetic images and corresponding lesion annotations. This metric is inspired by the study of label noise. Zhang *et al.* [12] revealed an intriguing phenomenon that the convergence time on false/noisy labels increases by a constant factor compared with that on true labels. Similar to this, aligning the well-transformed pixels (i.e., these pixels can be viewed as healthy ones) and the original lesion annotations is counter intuitive and hampers the fitting/convergence. To verify this, we assess the convergence speed by recording dice scores on the training data throughout the training process. The results on healthy images,

[1]To distinguish the segmentor in the downstream segmentation and the GVS, the downstream segmentor S_D is adopted in the downstream segmentation. similarly, in Sec. IV, the evaluation segmentor S_E denotes the segmentor using in the evaluation process.

[2]SI(GVS, a) is the abbreviation of the "Synthetic images (GVS, $a\%$)", which represents the synthetic images generated by the GVS that is trained on $a\%$ of training data.



(a) "Healthiness" [2] (b) Dice curve (lr=0.0001) (c) Dice curve (lr=0.1) (d) A-Dice

Fig. 5. (a) The "Healthiness" proposed in literature [2] fluctuates drastically at various epochs and runtimes, illustrating its instability. (b-c) Dice scores on the training data when $lr = 0.0001$ and $lr = 0.1$, respectively. Increasing the learning rate enlarges the gap of convergence speed. (d) The A-Dice scores evaluated on different synthetic images. The "SI(Method, a)" in the figure represents synthetic images generated by "Method" (e.g., GVS) which is trained on $a\%$ of training data. The proposed A-Dice accurately reflects the relative order for different kinds of images and fluctuates slightly.

the SI(GVS, 30) and SI(GVS, 100) are shown in Fig. 5(b). It is discovered that three evaluation segmentors finally attain similar dice values but have different convergence speeds. The evaluation segmentor S_E trained on healthy images achieves the fastest convergence speed, followed by the SI(GVS, 30) and SI(GVS, 100).

In the conference version, the $\mathbb{S}_{dice}$ was proposed to measure the healthiness by calculating the area under the dice curve. However, it may attain false results when synthetic images have similar healthy appearances since 1) the dice curves of synthetic images with different healthy appearances attain close convergence speed, which implies their lower discriminability, and 2) the performances at initial epochs are unstable due to the effects of random parameter initialization and batch selection, which results in the fluctuation of the $\mathbb{S}_{dice}$. To avoid false measurements, the first factor is considered, and the scheme that tries to *enhance the discriminability* between dice curves is proposed. Concretely, we further utilize another property of false/noisy labels. That is, the higher learning rate will suppress the memorization ability of the DNN and prevent it from fitting labels [13]. Improving the learning rate (i.e., from 0.0001 to 0.1) can obtain the dice curves presented in Fig. 5(c). We find that the gap of convergence speed between three types of images is significantly magnified. Thus, a new healthiness metric is proposed by evaluating the convergence speed of the evaluation segmentor S_E. The evaluation segmentor S_E is trained on $lr = 0.1$, and the dice values are recorded during the training process. Then, the A-Dice is calculated by averaging the dice values at multiple epochs, which is formulated as

$$\text{A-Dice} = \frac{1}{E}\sum_{e=1}^{E} dice_e, \tag{8}$$

where E denotes the total epochs and $dice_e$ represents the dice evaluated on the training data after the e-th epoch. *Note that a lower A-Dice denotes faster convergence speed and further represents more healthy appearances.* Furthermore, compared with $\mathbb{S}_{dice}$, the A-Dice replaces the summation operation with the mean operation to eliminate the impact of the training epoch.

We further emphasize that the proposed A-Dice is robust and stable. For the former one, any artifact distributionally different from the normal tissues can be quickly fitted to lesion annotations. As a result, it will be judged to be unhealthy due to the fast convergence. For the latter one, although the dice value at each epoch are unstable, the A-Dice can reduce the unreliability by averaging the dice values at multi epochs.

We also design four experiments to validate the reliability of the proposed A-Dice from multiple views. Firstly, the two artifacts mentioned in Fig. 4 are injected into the pathological images and then their A-Dice are evaluated. The green boxes in Fig. 5(d) show that their A-Dice values are little affected and even become larger. Thus, the proposed A-Dice *effectively identifies the pathology that slightly deviates from the true pathology but is apparently abnormal.* Secondly, the A-Dice values of ten types of images are presented in Fig. 5(d). The largest fluctuation range of the A-Dice is about 0.05 and is significantly less than the "Healthiness" [2] and the subjective assessment (rf. Table II), which *implies its good repeatability.* Thirdly, by comparing the relative relationship between different types of images, *the A-Dice accurately reflects the relative order of them.* That is, the A-Dice of SI(Method, a) is smaller than that of SI(Method, b) when $a > b$ and Method $=$ [GVS, ANT-GAN]. Fourthly, we find that *the A-Dice agrees with the subjective assessment to a certain extent.* That is, the PHS-GAN and ANT-GAN achieve close results for the A-Dice (0.607 and 0.618, rf. Table I), which also happens to the subjective healthiness metric (3.800 and 3.803, rf. Table II).

V. Experiments

A. Datasets

The proposed GVS is validated on two widely used public datasets: Multimodal Brain Tumor Segmentation Challenge 2019 dataset (BraTS19) [14]–[16] and Liver Tumor Segmentation Challenge dataset (LiTS) [17].

1) BraTS: The first validation dataset is the BraTS consisting of 259 GBM (i.e., glioblastoma) and 76 LGG (i.e., lower-grade glioma) volumes that have been skull-stripped, interpolated to an isotropic spacing of $1mm^3$ and co-registered to the same anatomical template. Each volume includes 4 modalities (i.e., T1, T2, T1c, and Flair), and each slice is 240×240. The T1 and T2 modalities of GBM are

utilized and split into training (234 volumes) and test sets (25 volumes). For each volume, the intensities are clipped to $[0, V_{99.5}]$, where $V_{99.5}$ is the 99.5% largest pixel value of the corresponding volume [42].

2) LiTS: We use the training data set of LiTS, which contains 131 CT scans of the liver acquired from 7 different clinical institutions. The resolution of the slice is 512×512. The dataset is divided into training (118 scans) and test sets (13 scans). Following Dou *et al.* [43], the image intensity values of all scans are truncated to the range of $[-200, 250]$ to remove the irrelevant details.

B. Implementation Details and Baselines

1) Implementation Details: The generator and the segmentor adopt the encoder-decoder [44] and U-Net [45] architectures, respectively. The proposed method is implemented in PyTorch. The models are trained with the Adam optimizer. The learning rate is set to 0.001, and the E is set to 20. The batch size is set to 8 for BraTS and 2 for LiTS. The λ is set to 10.0. The training is implemented using an NVIDIA TITAN XP GPU.

2) Baselines: The proposed method is compared with three approaches, VA-GAN [9], PHS-GAN [2], and ANT-GAN [5]. *The VA-GAN utilizes pathological images along with image-level annotations, whereas both the PHS-GAN and the ANT-GAN further utilize pathological images along with pixel-level annotations.* For the VA-GAN and PHS-GAN, we use the official code (i.e., VA-GAN[3] and PHS-GAN[4]) and train their architectures on our dataset. Besides, the ANT-GAN is implemented based on the code provided by the authors.

C. Other Evaluation Metrics

To comprehensively assess the effectiveness of our method, the synthetic images are evaluated objectively and subjectively.

1) Objective Metrics: The overall quality of pseudo-healthy images can be measured from two aspects, healthiness and subject identity. In Sec. IV, the proposed A-Dice has been described, which can properly assess the healthiness of pseudo-healthy images. This section further introduces the process to measure the subject identity, which can be expressed as calculating the visual similarity on the healthy regions. In practice, the MPSNR (masked PSNR) and MSSIM (masked SSIM) utilized to measure the subject identity are defined as.

$$\text{MPSNR} = \text{PSNR}[(1 - y_t) \odot \mathbf{G}(x_p), (1 - y_t) \odot x_p], \quad (9)$$

$$\text{MSSIM} = \text{SSIM}[(1 - y_t) \odot \mathbf{G}(x_p), (1 - y_t) \odot x_p], \quad (10)$$

where y_t is the lesion annotation while PSNR() and SSIM() denote Peak Signal-to-Noise Ratio and Multi-Scale Structural Similarity Index, respectively. Both PSNR and SSIM are adopted since they both are widely used to measure the similarity between two images and mainly differ on their degrees of sensitivity to image distortion [46].

[3]https://github.com/baumgach/vagan-code

[4]https://github.com/xiat0616/pseudo-healthy-synthesis

2) Subjective Metrics: The subjective metric is the gold standard to evaluate the quality of pseudo-healthy images. Here, we adopt the human evaluation method similar to the reference [2] is adopted, which is composed of the factors of healthiness and subject identity. The detailed process is presented as follows.

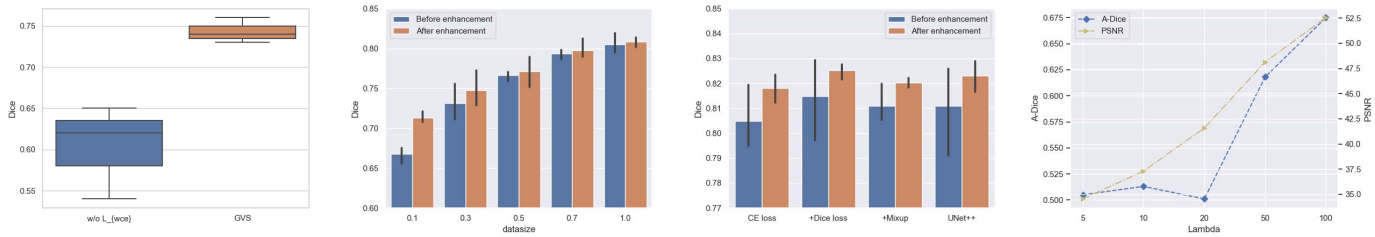
We randomly select 100 synthetic outputs from each comparison method and arrange them as follows. Except for the pathological input placed in the first position, the synthetic outputs of all comparison methods are randomly arranged in the next four positions. Finally, the last position is occupied by the lesion annotation to better convey the pathological information to raters. Meanwhile, the raters are blinded to the algorithm that generated each image. The subjective ratings are rated on a five-level scale: 5, 4, 3, 2, and 1. Three medical image analysis researchers are asked to independently score each synthetic image from the aspects of healthiness (i.e., the scores from 5 to 1 represent that the healthiness declines sequentially) and subject identity (i.e., the scores from 5 to 1 represent that the subject identity is erased sequentially). Moreover, the healthiness further can be judged from the degree of achieving harmony between healthy and pathological regions in the synthetic images. The criteria of subjective identity include the brightness, contrast, and the degree of keeping the detailed tissues.

D. Comparison With the State-of-the-Art Methods

1) Qualitative Results: The qualitative results are shown in Fig. 6. The effectiveness of all comparison methods is judged from the aspects of the subject identity and healthiness.

The healthiness can be judged by comparing whether pathological and normal regions are harmonious. If the pathological regions are in harmony with the normal regions in the synthetic images, such images are healthy. On the contrary, if the pathological regions can be easily distinguished from the normal ones in the synthetic images, such images are viewed as being "not healthy". The performance of the VA-GAN fluctuates substantially. Some of the synthetic images have promising performance (e.g., the first and third examples in Fig. 6). However, the majority of them can be easily distinguished from the healthy images due to poor reconstruction. The PHS-GAN and the ANT-GAN remove most lesions, but some artifacts still remain (rf. the first, second, fourth, and sixth samples in Fig. 6). Finally, the proposed GVS removes more lesions than the others and replaces the lesion regions with a relatively healthy-looking area.

The subject identity is determined by comparing the inputs and the synthetic images in terms of *structural details, brightness, and contrast*. Since the ANT-GAN, PHS-GAN, and the proposed GVS reconstruct the normal tissue well and have slight differences in reconstructions, we plot the difference maps between the inputs and the synthetic images to magnify their differences. Combining difference maps and labels, the proposed method shows higher quality reconstructions than the other methods due to preserving more details of the brain tissues. It should be noted that the cerebrospinal fluid has pixels with high intensities and is close to lesions. These



(a) (b) (c) (d)

Fig. 7. (a) The generalization ability of the segmentor in the GVS. Adding the $\mathcal{L}_{wce}$ effectively improves the generalization of the segmentor. (b) The variety of segmentation performance before and after enhancement. The proposed enhancement technique effectively improves the segmentation performance, especially in the low data regimes. (c) The variety of segmentation performance when combined with Dice loss, mixup, and UNet++. Combining these techniques and our enhancement method further improves the performance. (d) The sensitivity analysis for λ.

TABLE I

QUANTITATIVE RESULTS ON BRATS T2 MODALITY. WE REPORT THE AVERAGE VALUE AND STANDARD DEVIATION OF 3 TRIALS. GVS(a% DATA) REPRESENTS THE GVS TRAINED ON a% OF TRAINING DATA

Method	MPSNR ↑	MSSIM ↑	A-Dice ↓
Original images	-	-	0.736(±0.310)
VA-GAN [9]	21.89(±1.02)	0.742(±0.470)	-
PHS-GAN [2]	29.51(±0.47)	0.966(±0.025)	0.607(±0.056)
ANT-GAN [5]	28.75(±0.51)	0.963(±0.018)	0.618(±0.066)
GVS(30% data)	**38.79(±0.39)**	**0.995(±0.013)**	0.615(±0.040)
GVS w/o $\mathcal{L}_{wce}$	37.31(±0.34)	0.991(±0.012)	0.549(±0.045)
GVS	37.31(±0.34)	0.991(±0.012)	**0.512(±0.035)**

TABLE II

SUBJECTIVE RESULTS ON THE BRATS T2 MODALITY. WE REPORT THE AVERAGE VALUE AND STANDARD DEVIATION OF 3 RATERS

Method	"HEALTHINESS" ↑	"IDENTITY" ↑
VA-GAN [9]	2.103(±0.414)	1.943(±0.519)
PHS-GAN [2]	3.667(±0.824)	3.800(±0.568)
ANT-GAN [5]	3.707(±0.823)	3.803(±0.518)
Our GVS	**4.457(±0.330)**	**4.390(±0.491)**

regions are weakened to varying degrees in the other methods, whereas they are well-preserved in the proposed method. Overall, the VA-GAN cannot keep the subject identity and loses a part of the lesion region in some cases. The PHS-GAN and ANT-GAN preserve the brain region but lose some details. The proposed method achieves the best subject identity among all methods.

2) Objective Evaluation: The quantitative results on the BraTS dataset are shown in Table I. The results of MSSIM and MPSNR show that all methods significantly improve the subject identity compared with the VA-GAN. The reason is that the other methods utilize the pixel-level lesion annotations, which is important for keeping normal regions and transforming pathological regions. The proposed method further improves the visual similarity compared with the PHS-GAN and ANT-GAN. Otherwise, since the VA-GAN cannot reconstruct the normal tissues well (see Fig. 6 and Table I), its A-Dice value is meaningless and is not considered. The A-Dice values of the ANT-GAN and the PHS-GAN are 0.618 and 0.607. Compared to the original images, they decline by 0.118 and 0.129, respectively. Moreover, the A-Dice value of the proposed GVS is 0.512, with a decline of 0.224. The proposed method attains significant improvement compared with the existing methods. Even, trained on only 30% of the training data, the GVS achieves notable performance. Specifically, the A-Dice is close to the ANT-GAN and PHS-GAN while both the MSSIM and MPSNR significantly exceed them.

3) Subjective Assessment: We report the subjective results in Table II. The overall tendency of subjective results is similar to that of objective results. That is, the GVS shows the best performance in both "HEALTHINESS" and "IDENTITY" (i.e., subjective assessment for healthiness and identity. All capital for avoiding confusion.). Subsequently, the PHS-GAN and the ANT-GAN have similar results, while the VA-GAN occupies the bottom position. Particularly, the performance differences are magnified, and our GVS has better performance than the other methods in the aspects of both "HEALTHINESS" (GVS (4.457) vs ANT-GAN (3.707)) and "IDENTITY" (GVS (4.390) vs ANT-GAN (3.803)). Furthermore, our GVS achieves near-ideal performance ("HEALTHINESS": GVS (4.457) vs upper bound (5.0); "IDENTITY": GVS (4.390) vs upper bound (5.0)), which suggests the near-perfect performance of the GVS from the expert perspective.

E. Effectiveness of Difference-Aware Loss

Sec. III-B states that the segmentor in the basic GVS has a poor generalization ability due to the mismatch between synthetic images and lesion annotations. Thus, the generator will be misguided and further harm the synthetic quality. To verify this, we first evaluate the generalization ability of the segmentor before and after adding difference-aware loss. Specifically, the pathological images are sent to two segmentors trained by the GVS and the GVS w/o difference-aware loss, respectively, and then the corresponding dice scores are calculated. The results shown in Fig. 7(a) illuminate that the proposed GVS achieves a higher average dice score and lower variance compared to the GVS w/o $\mathcal{L}_{wce}$, which confirms that the difference-aware loss effectively improves the generalization ability of the segmentor.

Next, we further test and verify the effectiveness of difference-aware loss on healthiness and subject identity. As shown in Table I, the A-Dice has a significant improvement, while the MPSNR and MSSIM increase slightly,

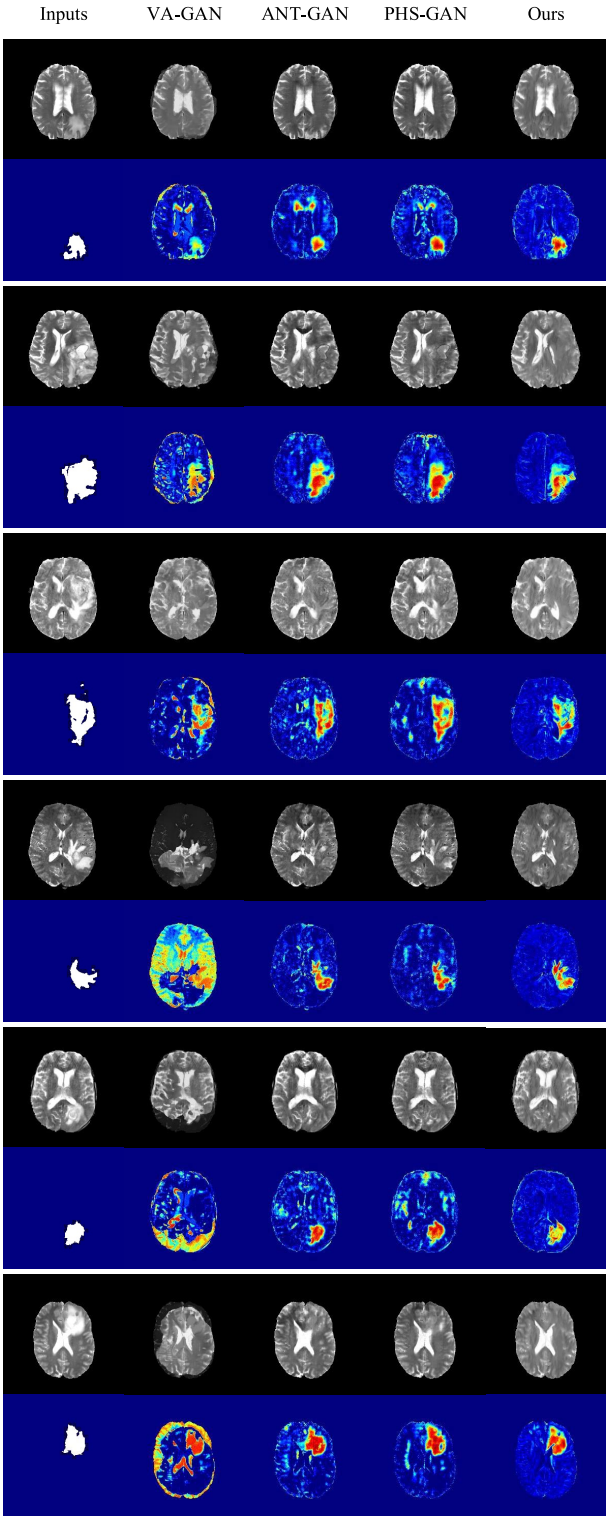


Fig. 6. Example results on the BraTS T2 modality. We plot six examples (blocks) from top to bottom. In each block, the first column shows the input and the lesion annotation, and the next four columns show the synthetic images and the difference maps generated by the VA-GAN, ANT-GAN, PHS-GAN and GVS, respectively.

which suggests a tight relationship between the generalization ability of the segmentor and the healthiness of synthetic images. Accordingly, improving the generalization ability of the segmentor is a plausible way to improve the healthiness of synthetic images.

F. Results on Contrast-Enhanced Brain Tumor Segmentation

The segmentation performance after enhancement is shown in Table III. It can be observed that under different values of α, the segmentation performances improve to different extents. To further explore the reason why the proposed enhancement technique can quantitatively improve the segmentation performance, four examples with lower contrast are shown in Fig. 8. Intuitively, it is easier to distinguish tumors after enhancement. Furthermore, the tumors in the enhanced images are detected more accurately. These results manifest that our method effectively simplifies the segmentation by enhancing the contrast between normal and pathological regions.

Next, we conduct the experiments to explore the improvements brought by the enhancement under different data sizes. The results are shown in Fig. 7(b), and we find that the improvements are more evident when the data size is smaller. We guess that this is because the network is easy to overfit when data size is small, and the enhancement explicitly regularizes the network at the image level so that better solutions are easier to be found.

Lastly, we verify the compatibility of the proposed enhancement by combining it with other techniques (e.g., Dice loss, mixup, and UNet++) that have proven effective for the segmentation. Specifically, the dice loss presented in V-Net [47] is an effective method to address the class imbalance problem existing in medical image segmentation. Mixup [48] is an effective data augmentation technique in the classification task. Recently, it has been introduced in medical image segmentation [49]. The U-Net++ [50] is a variant of U-Net based on nested and dense skip connections, and its effectiveness was verified on multiple medical image segmentation tasks. The results are shown in Fig. 7(c). The above-mentioned three techniques indeed improve the segmentation performance in brain tumor segmentation, and our enhancement technique further improves the segmentation performance on their basis. On the contrary, these three techniques also further improve the segmentation performance based on our enhancement technique.

G. Sensitivity Analysis of λ

The proposed GVS only contains one hyperparameter, λ, which is used to balance the power of the visual residual loss $\mathcal{L}_R$ and adversarial loss $\mathcal{L}_{s2}$ when training the generator G. Here, the sensitivity of GVS for different choices of λ is investigated and, the results are shown in Fig. 7(d). We discover that when $\lambda \in [5, 20]$, the healthiness varies slightly and is insensitive to the value of λ. Moreover, the healthiness and subject identity take up the opposite position. Better healthiness (lower A-Dice value) means worse subject identity and vice versa.

H. Extensive Results on Other Modalities

The proposed GVS is also evaluated on other modalities including CT and MR T1. The results of one CT dataset,

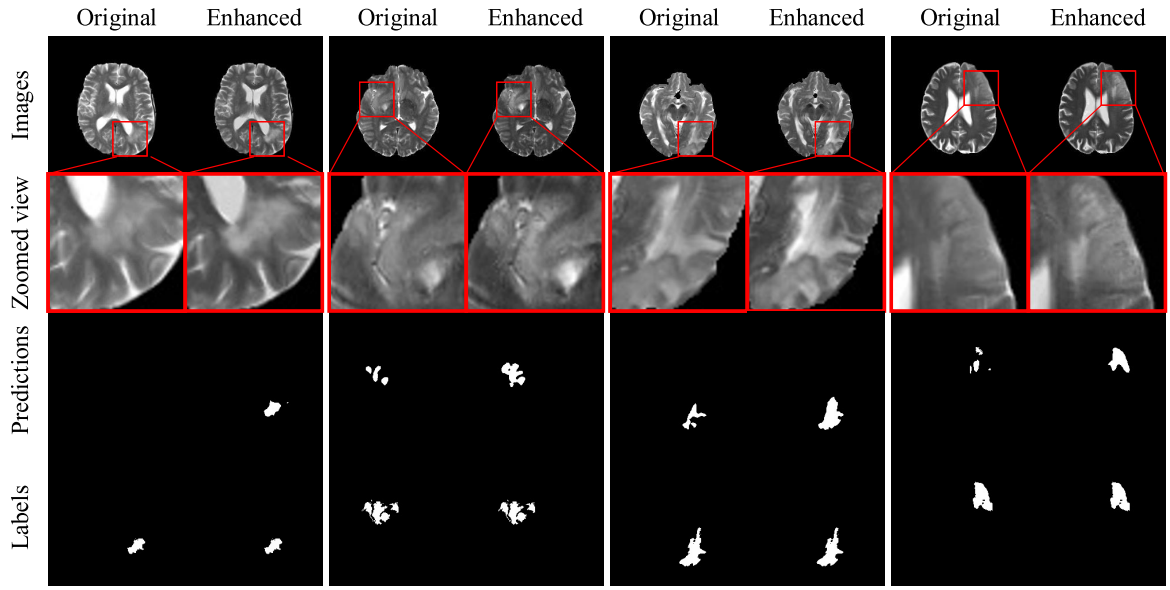


Fig. 8. Examples of segmentation predictions for enhanced results on the BraTS dataset. These results are generated when $\alpha = 1.0$.

TABLE III

COMPARISON OF THE SEGMENTATION PERFORMANCE WHEN $\alpha = [0.0, 0.3, 0.5, 0.7, 1.0]$. ALL RESULTS ARE AVERAGED OVER 3 TRIALS. NOTE THAT $\alpha = 0.0$ REPRESENTS THE IMAGES WITHOUT ENHANCEMENT, AND VALUES IN BRACKETS ARE DIFFERENCES OF SEGMENTATION PERFORMANCES BEFORE AND AFTER ENHANCEMENT. MAXIMUM IMPROVEMENT IS DENOTED IN **BOLD**, AND THE SECOND ONE IS DENOTED IN UNDERLINE

BraTS	α	0.0	0.3	0.5	0.7	1.0
	Dice	0.805	0.813(+0.008)	0.812(+0.007)	**0.818(+0.013)**	0.808(+0.003)
LiTS	α	0.0	0.3	0.5	0.7	1.0
	Dice	0.643	0.655(+0.012)	0.653(+0.010)	**0.666(+0.023)**	0.665(+0.022)

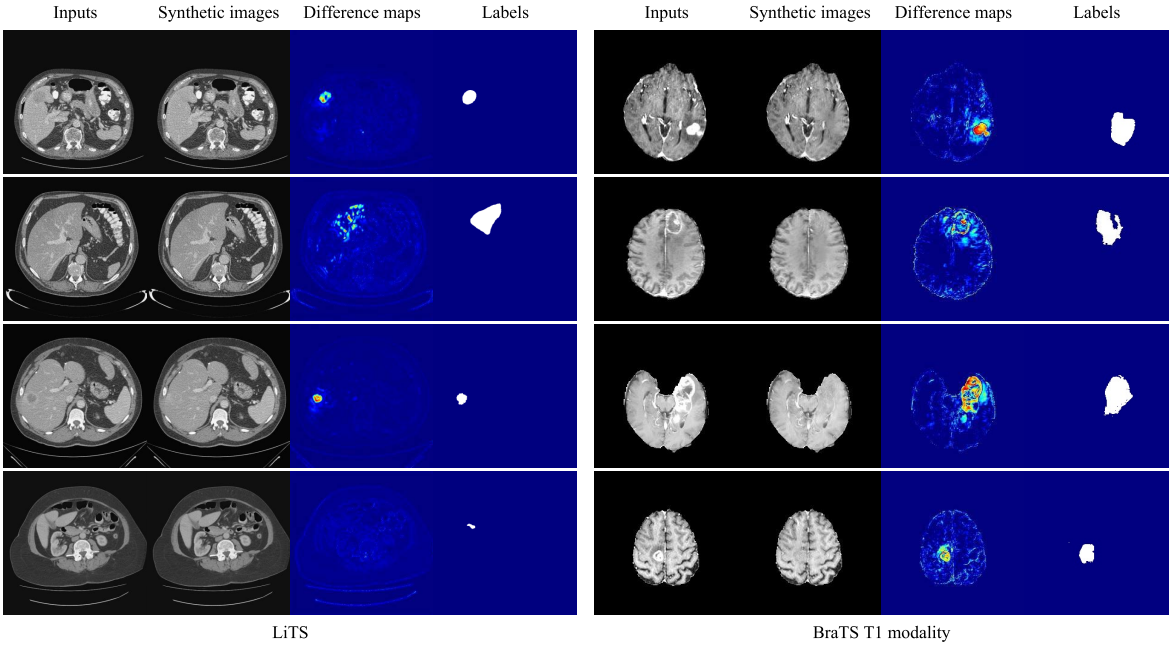


Fig. 9. Example results on the LiTS (left) and BraTS T1 modality (right). For each modality, we plot four examples from top to bottom. For each quaternion, we show the input, synthetic image, difference map, and label from left to right.

LiTS, are shown on the left side of Fig. 9. The top row shows that the proposed method effectively maintains the subject identity and transforms the high-contrast lesions into healthy-looking tissues well. However, the results corresponding to the images with lower contrast are not satisfactory, as shown in the bottom row. This is because the images in the LiTS dataset have higher resolution, and even more importantly, the lesions have smaller volumes and lower contrast. The right side of

Fig. 9 shows the results on the BraTS T1 modality, which are similar to that on the T2 modality. The results of MPSNR = 38.34, MSSIM = 0.994, and A-Dice = 0.501 also confirm the excellent performance of the proposed GVS.

The synthetic results are also applied to image enhancement to assist lesion segmentation, and the results are shown in Table III. It is observed that the performance improvement on the LiTS is more significant than that on the BraTS. We conjuncture that it's owing to the proposed enhancement technique facilitates the segmentation more significantly when it is applied to lower-contrast images. The phenomenon that the higher α results in more improvement further strengthens this view.

VI. Discussion and Conclusion

This work consist of three parts, involving an effective pseudo-healthy synthesis framework, a meaningful application, and a reliable evaluation metric.

The first one is an adversarial framework consisting of a generator and a segmentor proposed to cope with the problem of pseudo-healthy synthesis. The experiments show that our GVS achieves excellent performance compared to the state-of-the-art methods. Besides the results and the analysis in Sec. V-D, we further emphasize that the healthiness and the subject identity are competing (shown in Sec. V-G). A similar attribute, the contradiction between reconstruction and adversarial losses, is also verified in the GAN [51]. Hence, achieving superior performance in both healthiness and subject identity for our GVS is promising.

Fig. 6 and 9 show that our GVS fills pathological regions with diverse healthy-looking tissues, which implies the reduction of the mode collapse. The research [52], [53] demonstrated that the higher dimension will exacerbate the mode collapse and meanwhile need more training samples. In the GANs, one image is a sample, and its distribution is high-dimensional so that generating such distribution is easy to fall into collapse. Our GVS treats one pixel as a sample. The dimension of its distribution is significantly reduced. Thus, the mode collapse will be effectively alleviated.

Although our method achieves impressive results, it is confined to densely labeled annotations. In clinical applications, it is extremely difficult to collect huge amounts of accurate segmentation labels. Hence, it is necessary to relax the demand for accurate pixel-level annotations (e.g., semi-supervised learning, weakly-supervised learning) in the next step.

Sec. V-E reveals the close relationship between the power of the segmentor and the healthiness of synthetic images. It is found that strengthening the segmentor contributes to further improving the capacity of the GVS. One way to achieve this goal is to adopt more powerful architectures. Hence, we plan to upgrade the GVS to 3D structures to further improve the synthetic performance. Furthermore, it would also be worth exploring if the GVS could improve the segmentation performance of the segmentor in three dimensions.

Furthermore, we design the difference-aware loss to alleviate the poor generalization ability of the segmentor, which is verified to be effective for improving the healthiness of synthetic images (rf. Sec. V-E). However, this design still has room for improvement. For example, we do not consider that at first epochs the synthesis is imperfect and the difference maps are less meaningful. Hence, improving the difference-aware loss or designing a novel scheme to enhance the generalization of the segmentor is our next target.

The second part of the main work is using difference maps to enhance the contrast between normal tissues and lesions. The results in Sec. V-F and V-H illustrate that the proposed enhancement technique effectively alleviates the low contrast problem and improves the segmentation performance. Moreover, this technique is orthogonal to other techniques that also contribute to segmentation. More importantly, the segmentor trained on less training data benefits more from the enhancement is an interesting phenomenon. It is well known that training a model with small datasets is a critical and challenging topic in medical image segmentation [54]. The GVS provides an effective solution to deal with this problem and merits further exploration.

The third part is proposing a reliable metric, A-Dice, to measure the healthiness of synthetic images. The proposed metric also has the potential for wider applications. For example, we plan to use this metric to measure the abnormal information in the pathological images and explore the relationship between it and downstream tasks, such as tumor grading and survival prediction.

References

[1] C. Bowles *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *Proc. Int. Workshop Simulation Synth. Med. Imag.* New York, NY, USA: Springer, 2016, pp. 87–96.

[2] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101719.

[3] C. Bowles *et al.*, "Brain lesion segmentation through image synthesis and outlier detection," *NeuroImage, Clin.*, vol. 16, pp. 643–658, Jan. 2017.

[4] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: Coherent synthesis of subject-specific scans with data-driven regularization," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2013, pp. 606–613.

[5] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley, "An adversarial learning approach to medical image synthesis for lesion detection," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2303–2314, Aug. 2020.

[6] S. Andermatt, A. Horváth, S. Pezold, and P. Cattin, "Pathology segmentation using distributional differences to images of healthy origin," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2018, pp. 228–238.

[7] Y. Du, Q. Quan, H. Han, and S. K. Zhou, "Where is the disease? Semi-supervised pseudo-normality synthesis from an abnormal image," 2021, *arXiv:2106.15345*.

[8] Y. Tsunoda, M. Moribe, H. Orii, H. Kawano, and H. Maeda, "Pseudo-normal image synthesis from chest radiograph database for lung nodule detection," in *Advanced Intelligent Systems*. New York, NY, USA: Springer, 2014, pp. 147–155.

[9] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using Wasserstein GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8309–8319.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[11] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping," 2020, *arXiv:2002.11434*.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 2016, *arXiv:1611.03530*.

[13] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5552–5560.

[14] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, Sep. 2017, Art. no. 170117.

[15] B. H. Menze and *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2004.

[16] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[17] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[18] Y. Zhang *et al.*, "Generator versus segmentor: Pseudo-healthy synthesis," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 150–160.

[19] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders," 2018, *arXiv:1806.04972*.

[20] D. Sato *et al.*, "A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 105751P.

[21] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[22] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *Proc. Int. MICCAI Brainlesion Workshop*. New York, NY, USA: Springer, 2018, pp. 161–169.

[23] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2019, pp. 289–297.

[24] D. Zimmerer, S. A. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," 2018, *arXiv:1812.05941*.

[25] N. Pawlowski *et al.*, "Unsupervised lesion detection in brain CT using Bayesian convolutional autoencoders," Tech. Rep., 2018.

[26] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Scale-space autoencoders for unsupervised anomaly segmentation in brain MRI," in *Proc. MICCAI*, 2020, pp. 552–561.

[27] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab, "SteGANomaly: Inhibiting CycleGAN steganography for unsupervised anomaly detection in brain MRI," in *Proc. MICCAI*, 2020, pp. 718–727.

[28] B. Nguyen, A. Feldman, S. Bethapudi, A. Jennings, and C. G. Willcocks, "Unsupervised region-based anomaly detection in brain MRI with adversarial image inpainting," 2020, *arXiv:2010.01942*.

[29] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101952.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[31] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[32] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[34] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[36] M. Minderer, O. Bachem, N. Houlsby, and M. Tschannen, "Automatic shortcut removal for self-supervised representation learning," 2020, *arXiv:2002.08822*.

[37] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[38] N. Singh, L. Kaur, and K. Singh, "Histogram equalization techniques for enhancement of low radiance retinal images for early detection of diabetic retinopathy," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 3, pp. 736–745, Jun. 2019.

[39] Q. Wang, L. Chen, and D. Shen, "Fast histogram equalization for medical image enhancement," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2008, pp. 2217–2220.

[40] I. Frosio, G. Ferrigno, and N. A. Borghese, "Enhancing digital cephalic radiography with mixture models and local gamma correction," *IEEE Trans. Med. Imag.*, vol. 25, no. 1, pp. 113–121, Jan. 2006.

[41] M. Hamghalam, B. Lei, and T. Wang, "High tissue contrast MRI synthesis using multi-stage attention-GAN for glioma segmentation," 2020, *arXiv:2006.05030*.

[42] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[43] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[44] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 234–241.

[46] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[48] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[49] Z. Eaton-Rosen, F. Bragman, S. Ourselin, and M. J. Cardoso, "Improving data augmentation for medical image segmentation," Tech. Rep., 2018.

[50] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. New York, NY, USA: Springer, 2018, pp. 3–11.

[51] S. Ramasinghe, K. Ranasinghe, S. Khan, N. Barnes, and S. Gould, "Conditional generative modeling via learning the latent space," 2020, *arXiv:2010.03132*.

[52] N.-T. Tran, T.-A. Bui, and N.-M. Cheung, "Improving GAN with neighbors embedding and gradient matching," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5191–5198.

[53] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[54] P. Zhang, Y. Zhong, Y. Deng, X. Tang, and X. Li, "A survey on deep learning of small sample in biomedical image analysis," 2019, *arXiv:1908.00473*.