IEICE TRANS. INF. & SYST., VOL.E106–D, NO.5 MAY 2023

PAPER Special Section on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications

Image-to-Image Translation for Data Augmentation on Multimodal Medical Images

Yue PENG[†a)], ***Member***, **Zuqiang MENG**[†b)], ***and*** **Lina YANG**[†], ***Nonmembers***

SUMMARY Medical images play an important role in medical diagnosis. However, acquiring a large number of datasets with annotations is still a difficult task in the medical field. For this reason, research in the field of image-to-image translation is combined with computer-aided diagnosis, and data augmentation methods based on generative adversarial networks are applied to medical images. In this paper, we try to perform data augmentation on unimodal data. The designed StarGAN V2 based network has high performance in augmenting the dataset using a small number of original images, and the augmented data is expanded from unimodal data to multimodal medical images, and this multimodal medical image data can be applied to the segmentation task with some improvement in the segmentation results. Our experiments demonstrate that the generated multimodal medical image data can improve the performance of glioma segmentation.
key words: *data augmentation, image-to-image translation, gliomas*

1. Introduction

Many problems in image processing, computer graphics, and computer vision can be posed as "translating" an input image into a corresponding output image [1]. Image-to-image translation aims to learn a mapping that can transfer an image from a source domain to a target domain, while maintaining the main representations of the input image [2]. Applications include image super-resolution [3], domain adaptation [4], image colorization [5], style transfer and data augmentation. Generative Adversarial Networks (GAN) work by specifying a high-level goal to make the output indistinguishable from reality, and then automatically learning a loss function suitable for satisfying this goal [6].

GAN proposed a minimax game between two Neural Networks Generator generates samples, discriminator identifies the source of samples [7]. Conditional generative adversarial networks can achieve image-to-image translation by using paired training data [8]. Pix2Pix is a generic image-to-image translation algorithm using CGAN [1], [8]. It can produce reasonable results on a wide variety of problems. Given a training set which contains pairs of related images, Pix2Pix learns how to convert an image of one type into an image of another type, or vice versa. To solve the problem that Pix2Pix is only applicable to pairs images, Cycle-Consistent GAN [9] were proposed, It use cycle consistency loss to store key attributes between the input and transformed images to train unpaired images. Research shows that Cycle-Consistent GAN have better performance than Pix2Pix. However, either Pix2Pix or Cycle-Consistent GAN only solves the problem of transforming from one domain to another, in the field of image translation. For example, having k domains, these methods require to train $k(k-1)$ generators to handle translations between each and every domain, limiting their practical usage [10].

To address the scalability, several studies have proposed a unified framework [11], StarGAN solves the problem of image translation between multi-domain just by adding control information of one domain (Choi et al. 2018). In the study of image-to-image translation, StarGAN has the advantage that only a single generator needs to be learned and has a more significant translation effect. However, StarGAN still learns a deterministic mapping per each the data distribution. This limitation comes from the fact that each domain is indicated by a predetermined label, and StarGAN V2 is a scalable approach that can generate diverse images across multiple domains [10] (Choi et al. 2020). In particular, StarGAN V2 does not require a predetermined label, and only needs to input the image of the original domain and the specified reference image of the target domain to complete the image translation.

It is widely known that sufficient data is critical to success when training deep learning models for computer vision. While traditional data augmentation schemes (e.g., crop, rotation, flip, and translation) can mitigate some of these issues, those augmented images have a similar distribution to the original images, leading to limited performance improvement [12]. Also, the diversity that those modifications of the images can bring is relatively small. Motivated by the GAN, researchers try to add synthetic samples to the training process. GAN-based data augmentation can improve performance by filling the distribution that uncovered by origin images. Since it can generate new but realistic images, it can achieve outstanding performance in medical image analysis [13], [14].

Gliomas are considered as an alarming and increasing brain tumors which seriously affect human mortality rate [15]. These tumors are characterized into two core types-Low Grade Gliomas (LGG) and High Grade Gliomas (HGG) [16]. LGG tumors can be classified as benign and malignant tumors, which grow very slowly in brain cells, therefore, patients can survive for several years; HGG tumors belong to only malignant tumors, which grow faster in brain cells, therefore, patients have a life expectancy of no more than 2 years [17]. The overall assessment shows that

Manuscript received September 13, 2021.
Manuscript revised January 5, 2022.
Manuscript publicized March 1, 2022.
[†]The authors are with Guangxi University, China.
a) E-mail: py20121121@163.conm
b) E-mail: zqmeng@126.com (Correnponding author)
DOI: 10.1587/transinf.2022DLP0008

the survival of LGG and HGG patients cannot survive more than 14 months [18], and the manual process of diagnosing these tumors is laborious and time-consuming. Hence, in clinical practice, MRI is useful in the assessment of gliomas because it can provide important information. For this reason, MRI is useful in assessing gliomas in clinical practice because it provides important information about the tumor area. Compared to CT and PET images, MRI provides soft tissue contrast. In MRI, tumors are classified into four modalities of information. These types are FLAIR, T1, T1c and T2, as shown in Fig. 1. In general, multi-modality data can lead to better performance results compared to a single-modality based approach, as different imaging methods can capture more information about the tumor [19], [20]. In order to better extract the details of medical images and save more image energy, [21] proposes a novel MST-based method. This novel fusion method enhances the computational efficiency and fusion performance, as well as improves the visual perception of images. [22] proposes a CNN-based medical image fusion method, the proposed fusion algorithm can effectively preserve the detailed structure information of source images and achieve good human visual effects. In addition, the effect of CNN on image processing, [23] proposes an image fusion-based algorithm to enhance the performance and robustness of image dehazing. Based on a set of gamma-corrected underexposed images, pixelwise weight maps are constructed by analyzing both global and local exposedness to guide the fusion process. [24] verifies that the dehazed images obtained by the patch based MEF algorithm always meet the requirements of intensity decrease. [25] proposed an automated segmentation technique followed by self-driven post-processing operations to detect cancerous cells effectively. [26] attempt to provide a detailed discussion on different types of adversarial attacks with various threat models and also elaborate on the efficiency and challenges of recent countermeasures against them.

Based on the above research, we know that multimodal medical images are composed of different modal data and represent different brain features, which is a necessary prerequisite for the realization of multi-domain transformation. In the field of computer-aided diagnosis, multimodal medical images can be used for image segmentation to assist doctors in diagnosis and treatment. Therefore, data augmentation of multimodal medical images is of great significance. In this paper, we propose to use an improved StarGAN V2 network model for data augmentation to improve glioma segmentation.

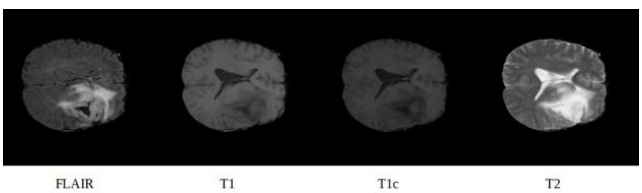

Fig. 1 FLAIR, T1, T1c, T2 sequence of original image slice.

The main contributions of this paper are:

· This paper proposes an image-to-image translation framework based on StarGAN V2, and attempts to combine the framework with medical images, which has a positive impact on the model even in the case of insufficient data.

· In the case of using only single modal data, the multimodal medical image data generated after data augmentation in the designed framework can be used for medical image segmentation through experiments in this paper.

· We add an attention module that reduces the complexity of the model and significantly improves the performance of the model. Experimental results show that the proposed design can augment the multimodal brain tumor dataset and improve the tumor segmentation performance of multimodal data even when the amount of data is insufficient.

2. Related Work

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GAN) are powerful generative models, which have achieved impressive results in many computer vision tasks [12]. A typical generative adversarial model consists of two modules: a discriminator and a generator (Fig. 2). The main role of Generator (G) is to generate images. First, G will receive random noise and use the random noise to generate a pseudo-image labeled G(z). Discriminator (D) is responsible for distinguishing whether the input image is a real image or not. If $D(x) = 1$, it means that this image is distinguished as Real. if $D(x) = 0$, it is the generated Fake image. The main role of D in GAN is to distinguish whether the input image is a real image or not and provide feedback.

To make the generated images indistinguishable from real images, the loss function is constructed in GAN [27]. Firstly, $G(z)$ is generated based on random noise, and secondly, $D(x)$ is trained to distinguish the real or fake images, and the two steps are alternated to form a dynamic game. The formula for GAN is as follows:

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

Where x as the input data, which is represented as the real

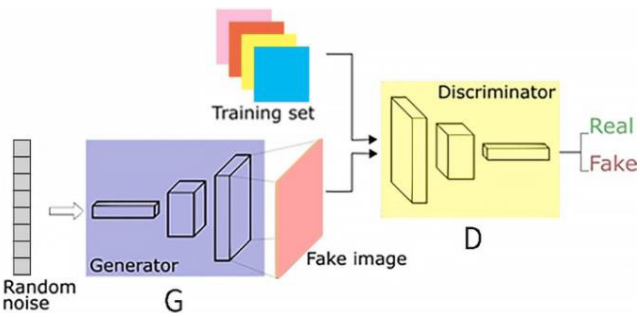

Fig. 2 Typical framework of GAN

Fig. 3 Model of StarGAN

image of the model input, and z indicates the random noise. $G(z)$ is represented as the new image output from the generative G, and D represents the calculated mistake between the Fake image and the Real image, represented by $D(x)$. $D(G(z))$ is the parameter representing the probability that D distinguishes whether the image generated by G is a real image or a fake image.

We can see that for G, it's better for $D(G(z))$ to be as big as possible, which means that $V(D,G)$ is as small as possible. Therefore, it is not difficult to understand that the minimum for G is min_G. The larger the value of $D(x)$, the smaller the value of $D(G(x))$, indicating that D is more powerful. At this time $V(D,G)$ will become bigger. In other words, Eq. (1) represents finding the maximum value max_D of D. P_{data} means the distribution of the original data and P_g stands for the distribution of the generated data. Equation (1) shows that when $D(x)$ and $G(z)$ are trained optimally, there are $P_{data} = P_g$ and $D(x) = 0.5$ for P_{data} and P_g, which means that the data distribution of the generated image matches that of the original image. It means that the quality of the generated images is sufficiently great to be indistinguishable by the discriminator. The poor efficiency in image-to-image translation is due to the fact that when dealing with image translation tasks between multiple domains, different models must be trained for each domain. And the impact on the quality of the generated images is caused by not fully utilizing the training data. StarGAN [28] was proposed to solve these problems by implementing the translation between multiple domains in the work of image-to-image translation using a single generator (Fig. 3).

2.2 StarGAN Works

The contributions of *StarGAN* are summarized as follows:

- To guide the generation of each domain image, the target domain information y (Fig. 4) is added as input in the model G. At this point, the adversarial loss $\mathcal{L}_{adv}$ can be obtained as follows:

$$\mathcal{L}_{adv} = E_x[logD_{src}(x)] + E_{x,c}[\log(1 - D_{src}(G(x,y)))] \quad (2)$$

  Where $D_{src}(x)$ denotes the probability distribution of the data given by discriminator D.
- In addition to the ability to distinguish whether the image is a real image, the discriminator D needs to learn how to classify the image into its corresponding domain. For the same input image, generator G needs to ensure that it can generate different images from different target domains, thus obtaining the domain classification loss $\mathcal{L}^D_{cls}$ and $\mathcal{L}^G_{cls}$:

$$\mathcal{L}^D_{cls} = E_{x,y'}[-logD_{cls}(y' \mid x)] \quad (3)$$

$$\mathcal{L}^G_{cls} = E_{x,y}[-logD_{cls}(y \mid G(x,y))] \quad (4)$$

  Where $D_{cls}(y' \mid x)$ represents the probability distribution over the original domain y' as calculated by D.
- Image reconstruction is to ensure the preservation of image content in the process of image transformation. Firstly, the image is converted from the original domain to the target domain, and then the image of the transformed target domain is converted back to the original domain. In order to ensure the consistency of the transformed image with the original image, the cyclic consistency loss $\mathcal{L}_{rec}$ is proposed:

$$\mathcal{L}_{rec} = E_{x,y,y'}[\|x - G(G(x,y),y')\|_1] \quad (5)$$

  where x is the input image, y' is the original domain, and y is the target domain. StarGAN uses (x, y) as input to the model G, when training the model G to generate the fake image x', and then inputs (x', y') to the model G and uses it to reconstruct the image x. StarGAN uses (x, y') and (x', y) as inputs to the discriminator when training the model D.

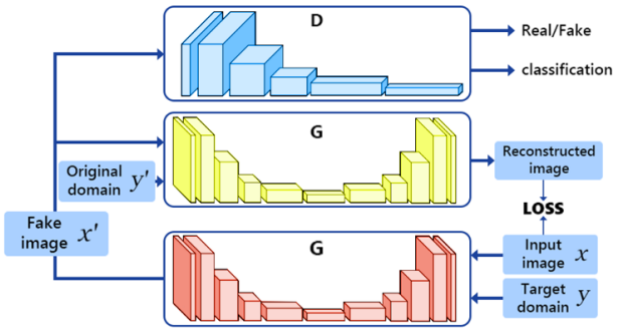

Fig. 4 Framework of StarGAN.

In Fig. 4, the job of discriminator D is to judge whether the image is true or false and classify the image into its corresponding domain. In general, D needs to learn how to classify images into their corresponding domains. g needs to ensure that it can generate different images from different target domains.

3. Network Construction

3.1 Network Structure

In this section, we describe StarGAN V2, which is an improved model of StarGAN with the original advantages. This section describes the theoretical analysis applied to multimodal medical image data augmentation and the network construction.

The structure of the ACGAN [24] discriminator is referenced in StarGAN, a typical GAN model has only random noise z as an input variable. ACGAN is different in that it has an additional classification variable. Typical GAN model outputs only the judgment of image authenticity, while ACGAN model adds classification judgment on this basis.

ACGAN not only adds image classification to distinguish true or false images but also improves the original discriminator D to distinguish between true and false images in different domains, so that the generator can translate the whole structure.

The W distance proposed by Wasserstein GAN [29] (WGAN) is appropriate for the training of GAN, and according to the theory analysis, the loss function of WGAN is the W distance from one distribution to the other. The W is calculated by:

$$W(P_r, P_f) = inf_{\gamma \in \prod(P_r, P_f)} E_{(x,y)\sim\gamma}[\|x - y\|] \tag{6}$$

Where $\prod(P_r, P_f)$ is the joint distribution function of all marginal distributions as Pr and Pf.

WGAN-GP [30] adds the concept of Gradient Penalty (GP) to WGAN, and GP is the gradient term of the real image that evaluates the difference of the generated image. The calculation formula is as follows:

$$\mathrm{R}(\psi) = \frac{\gamma}{2} E_{\hat{x}}(\|\nabla_x D_\psi(\hat{x})\| - g_0)^2 \tag{7}$$

Where $\hat{x}$ is for a random spot uniformly sampled on a line segment between two random points.

Based on this, R1 norm with Adaptive Instance Normalization (*AdaIN*) [31] is introduced to increase the stability. The calculation formula is as follows:

$$\mathrm{R}_1(\psi) = \frac{\gamma}{2} E_{pD(x)}[\|\nabla D_\psi(x)\|]^2 \tag{8}$$

$$AdaIN(z, \gamma, \beta) = \gamma\left(\frac{z - \mu(z)}{\sigma(z)}\right) + \beta \tag{9}$$

where z, μ represents the activation function generated by the previous convolutional layer, σ is the mean and standard deviation based on the channel, and γ, β are the parameters generated by the *MLP*.

For scalability, StarGAN uses a mapping network to learn each domain, which achieves high scalability of the model. Since StarGAN proposes to use predefined labels, the issue of diversity is not fully considered. A mapping network structure is added to the style encoder on this basis, and the output is the style code of the target model. The first step learns to convert random Gaussian noise into style codes, and the second step learns to extract style codes from a given reference image, thus achieving diversity and scalability.

Assuming x is the input image described in the previous section (Fig. 4), X represents the set of input images, and Y denotes the set of target domains. Given an image $x \in X$ and $y \in Y$ in the target domain, the aim is to train a single generator G that generates images corresponding to image x that belong in y. Domain-specific style vectors are generated in the learning style space of each domain, and G is trained to reflect the generated medical image modal vectors. We show the parts related to image generation in the example given in Fig. 5. The parts of StarGAN V2 related to image generation are shown in Fig. 5. They include the ResNet-like [32] (ResBlk) encoder (The blue parts in Fig. 5 (a) represent the ResBlk module), the decoder with the AdaResBlk module (yellow part in Fig. 5 (a) and (b)), and a set of condition-related style information encoders (green section in Fig. 5 (a)) with a shared header layer (purple section in Fig. 5 (a)).

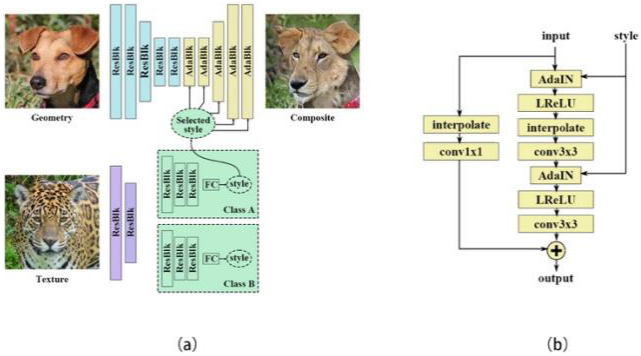

Fig. 5 In StarGAN V2, (a) the parts related to image generation. (b) AdaResBlk module.

G translates the input image x into the output image $G(x, s)$, reflecting the domain-specific style code provided by the mapping network F or style encoder marked as s. The style is represented in the normalization layer using affine transform parameters and the residual block uses the AdaIN layer with the parameters dynamically generated from s by the multi-layer perceptron (*MLP*) [31]. Adding s to G. We show that the network of generators is constructed as below (Table 1). We used Average pooling (AvgPool) because we need to retain the background information and also the AvgPool extracts smoother features. The role of Conv1×1 is to up-dimension/down-dimension the number of channels.

Given a latent code z and a domain y in the mapping network F, F will generate a style code s.

$$s = F_y(z) \tag{10}$$

Where $F_y(z)$ represents the output of F according to y. F consists of an *MLP* with multiple output branches, which can provide s. In addition, F can generate different s by randomly sampling $y \in Y$. Thus allowing F to learn the style representation of each modality efficiently and effectively in the task of generating multimodal medical images. We show the construction of the mapping network as below (Table 2). ReLU denotes the Rectified Linear Unit [33].

Given an image x and its correlation domain y in the style encoder, the encoder E will extract the style code of x.

$$s = E_y(x) \tag{11}$$

Where $E_y(x)$ represents the output of E in relation to y. Sim-

Table 1 Framework of generator

Layer	Resample	Norm	Output Shape
Image X	-	-	256×256×1
Conv1×1	-	-	256 × 256 × 64
ResBlk	AvgPool	IN	128×128× 128
ResBlk	AvgPool	IN	64× 64 × 256
ResBlk	AvgPool	IN	32 × 32 × 512
ResBlk	AvgPool	IN	16 × 16 × 512
ResBlk	-	IN	16 × 16 × 512
ResBlk	-	IN	16 × 16 × 512
ResBlk	-	AdaIN	16 × 16 × 512
ResBlk	-	AdaIN	16 × 16 × 512
ResBlk	Upsample	AdaIN	32 × 32 × 512
ResBlk	Upsample	AdaIN	64× 64 × 256
ResBlk	Upsample	AdaIN	128×128× 128
ResBlk	Upsample	AdaIN	256 × 256 × 64
Conv1×1	-	-	256 × 256 × 1

Table 2 Framework of mapping network

Type	Layer	Activation	Output Shape
Shared	Latent z	-	16
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
UnShared	Linear	ReLU	512
UnShared	Linear	ReLU	512
UnShared	Linear	ReLU	512
UnShared	Linear	-	64

ilar to F, E can generate different s with reference to different modal images. this allows the output of G to be combined with an image that reflects a different modal image x with s.

The discriminator D is a multi-task discriminator [34], which consists of multiple branches. A binary classification is learned with each branch D_y, which is used to distinguish whether x is a real image in y or a false image generated by $G(x, s)$. We show the network construction of the discriminator as below (Table 3). LReLU indicates the Leaky Rectified Linear Unit [35].

We will explain StarGAN V2 in detail, as shown in

Table 3 Framework of discriminator

Layer	Resample	Norm	Output Shape
Image X	-	-	256×256×1
Conv1×1	-	-	256 × 256 × 64
ResBlk	AvgPool	-	128×128× 128
ResBlk	AvgPool	-	64× 64 × 256
ResBlk	AvgPool	-	32 × 32 × 512
ResBlk	AvgPool	-	16 × 16 × 512
ResBlk	AvgPool	-	8 × 8 × 512
ResBlk	AvgPool	-	4 × 4 × 512
LReLU	-	-	4 × 4 × 512
Conv4× 4	-	-	1 × 1 × 512
LReLU	-	-	1 × 1 × 512
Reshape	-	-	512
Linear * K	-	-	D * K

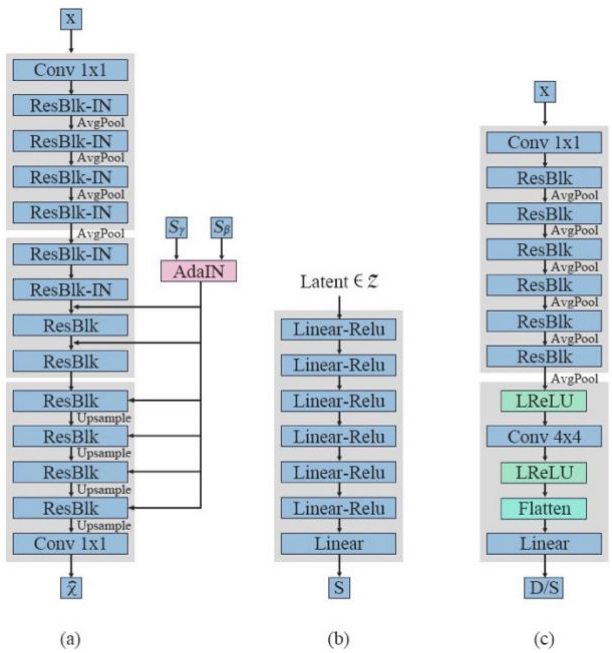

Fig. 6 Design diagram of the concrete implementation model of starGAN V2. (a) Generator. (b) Mapping network. (c) Style encoder/discriminator.

Fig. 6 (b), the input of the mapping network is a priori z. The mapping network learns the style representation of the target domain image. That is, the source domain image x provides the general content information of the source domain image, while the style encoding provides the style representation of the target domain, and this part can guide the source domain image x to transform to the image under the target domain according to the corresponding style. Since each sample is

different and the obtained style representation is also different, different style images under the target domain can be generated from a fixed x according to different styles.

The style encoder is not simply used to optimize the mapping network, but its other major role is to guide the style of the target domain images in the application phase, that is, if you want to specify the generator to generate and request the style of a certain image y in the target domain, what should be sent in at this time is the output of the style encoder at this time. So that the generated image is the specified style, here it can be understood that the style encoder in the test phase is an online label generator, used to specify the generator (Fig. 6 (a)) to follow what style to transform.

The style encoding combined with the source domain input image x can be fed to the generator, which outputs the transformed target domain image, while the discriminator (Fig. 6 (c)) is used to distinguish whether the generated target domain image is truly derived from the real target domain.

In this paper for the construction of a framework based on StarGAN V2 for multimodal medical image data augmentation. The generator consists of four down-sampling blocks, four intermediate blocks and four up-sampling blocks, and each layer has ResBlk.

During training, a sample of latent code $z \in Z$ and a target domain $\tilde{y} \in Y$ are randomly selected and a target style code is generated.

$$\tilde{s} = F_{\tilde{y}}(z) \tag{12}$$

x and $\tilde{s}$ are used as inputs to generate the output image $G(x, \tilde{s})$, where the loss function is.

$$\mathcal{L}_{adv} = E_{x,y}[logD_y(x)] + E_{x,\tilde{y},z}[\log(1 - D_{\tilde{y}}(G(x, \tilde{s})))] \tag{13}$$

The generator G generates $\tilde{s}$ with a loss function at this point.

$$\mathcal{L}_{sty} = E_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x, \tilde{s}))\|_1] \tag{14}$$

The paper [36] uses more than one encoder to learn the mapping from an image to its underlying code. The paper [10] trains an encoder E to encourage diverse outputs from multiple domains. At the time of training in this paper, the learned encoder E allows the style of the G mapping reference image to transform the input image.

To further enable the generator G to produce diverse images, the regularization of G for diversity-sensitive losses is required [16].

$$\mathcal{L}_{ds} = E_{x,\tilde{y},z_1,z_2}[\|G(x, \tilde{s}_1) - G(x, \tilde{s}_2)\|_1] \tag{15}$$

where the target style codes $\tilde{s}_1$ and $\tilde{s}_2$ are generated by F conditional on two random latent codes z_1 and z_2.

In order to ensure that the generated image $G(x, \tilde{s})$ ensures the domain-invariant properties of its input image x, the cycle consistency loss will be used [28]

$$\mathcal{L}_{cyc} = E_{x,y,\tilde{y},z}[\|x - G(G(x, \tilde{s}), \hat{s})\|_1] \tag{16}$$

Where $\hat{s} = Ey(x)$ is the estimated style code of the input image x and y is the original domain of x. By stimulating the generate G to reconstruct the input image x with the estimated style code $\hat{s}$ G is trained to ensure that the style of x is changed while the original features of x are preserved.

In summary, the full objective function is as in the public

$$min_{G,F,E}\, max_D \quad \mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc} \tag{17}$$

where λ_{sty}, λ_{ds} and λ_{cyc} are the hyperparameters of each term.

3.2 Channel Attention Module

The residual blocks build the entire model, and to enhance the model's ability to capture hierarchical patterns and thus improve the quality of the image representation, we use an improved Efficient Channel Attention (ECA) module [37].

Suppose the output of a convolution block is $\mu \in R^{w \times H \times C}$ where W, H, and C are the width, height, and channel size (i.e., the number of filters). Therefore, the weights of the channels in the SE block [38] can be calculated as

$$\upsilon = \sigma(f\{W_1, W_2\}(g(\mu))) \tag{18}$$

Where $g(\mu) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H}\mu_{ij}$ is channel-wise global average pooling (GAP) and σ is a Sigmoid function. Here we assume that $\theta = g(\mu)$, $f\{W_1, W_2\}$ takes the form

$$f\{W_1, W_2\}(\theta) = W_2\, ReLU(W_1\theta) \tag{19}$$

where $ReLU$ denotes the Rectified Linear Unit [33]. To avoid high model complexity, size of W_1 is set to $C \times \left(\frac{C}{r}\right)$, and W_2 is set to $\left(\frac{C}{r}\right) \times C$. It can be seen from Eq. (19). that $f\{W_1, W_2\}$ involves all parameters of the attention block of the channel. Although the dimensionality reduction in Eq. (19) reduces the complexity of the model, it destroys the direct correspondence between the channels and their weights. We can see the difference in Fig. 7.

By analyzing the impact of channel dimension reclassification and cross-channel interactions on channel attention learning above, the ECA module is used, and ECA does the following: the first step is to obtain a $1\times1\times D$ vector by GAP. The second step is to complete the information interaction across channels by 1D-Conv. The size of the convolution kernel of the 1D convolution is self-adapted by a function

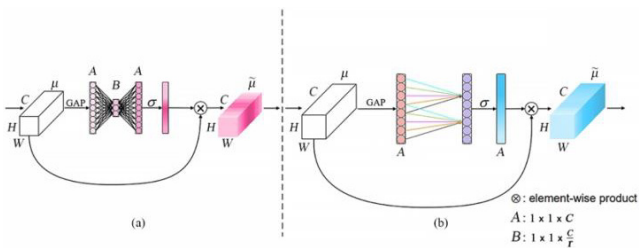

Fig. 7 SE module and ECA module

that allows more cross channel interactions for layers with a larger number of channels. The size of the adaptive convolution kernel is calculated as:

$$k = \varphi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right| \tag{20}$$

Where γ and b are set to 2 and 1, respectively, by mapping φ, the high-dimensional channels have a longer range of interactions, while the low-dimensional channels experience a shorter range of interactions by using a nonlinear mapping.

In Fig. 7, (a) represents the SE block, using two FC layers to calculate the weights. (b) represents the ECA block, which generates the channel weights by performing a fast one-dimensional convolution of size k, where k is determined adaptively as a function of the channel dimension C.

4. Experiment Results and Analysis

In this section, we will verify whether image-to-image translation enables data augmentation of multimodal medical images and verify whether the augmented data can reduce the dependence on the dataset. UNet [39] was tested using raw and extended multimodal medical image datasets (We label it as model$_1$ shown in Table 8), and 2018 Brain Tumor Segmentation Challenge Winner Method (We label it as model$_2$ shown in Table 8) [40], respectively.

The training data for this paper were selected from the BRATS [41] dataset, which has 285 cases with four modalities of information (FLAIR, T1, T1c, T2). The data were randomly dropped in the preprocessing step, as if each case had a modality information. Since the complete modality dataset is difficult to obtain in multimodal medical images, because in addition to data collection, alignment needs to be done. we use single-modal data with image size of $240 \times 240 \times 155$ for each modality as the training data for the theoretical study in this paper. All training processes were performed in an NVIDIA GeForce GTX 1080 Ti GPU.

In the data augmentation experiment. We use MUNIT [2] and original StarGAN V2 as our baselines, which learn multimodal mappings between two domains. In the medical image segmentation experiments, we compare the original data with the multimodal medical data after data augmentation in this paper in the same segmentation experiments, and there is no difference between the two except for the source of the dataset. The segmentation experiment is to verify whether the augmented data can be effective, so the validation experiments are conducted under model$_1$ and model$_2$ with the same experimental environment, respectively.

We evaluate Generated data on diverse image synthesis from two perspectives: latent-guided synthesis and reference-guided synthesis [10]. Figure 8 we explicitly label the original data as well as the reference images. Reference-guided image synthesis results on BRATS, and in Fig. 9 shows us the training process, labeled as FLAIR indicates that we use the existing modality FLAIR in the GAN to generate the training process images of other modalities, T1, T1c, T2 in the same way, The source and reference images in the first row and the first column are real images, while the rest are images generated by ours model, and it learns to transform a source image reflecting the style of a given ref-

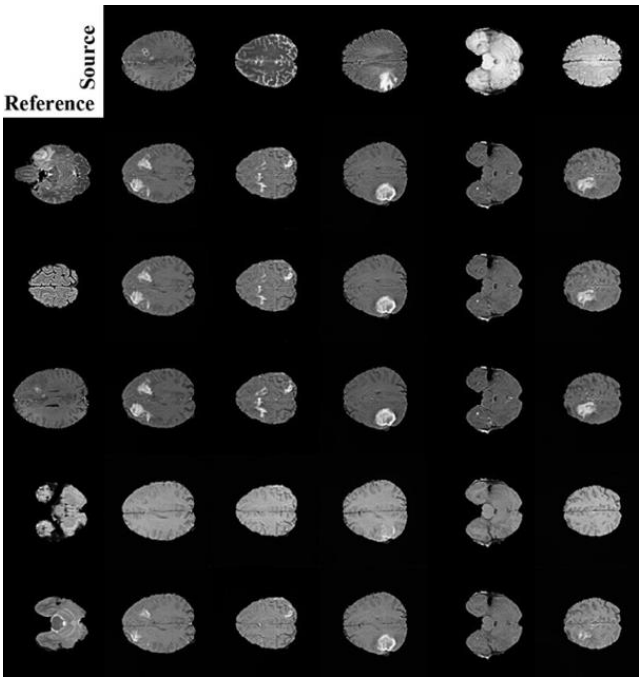

Fig. 8 Reference-guided image synthesis results on BRATS

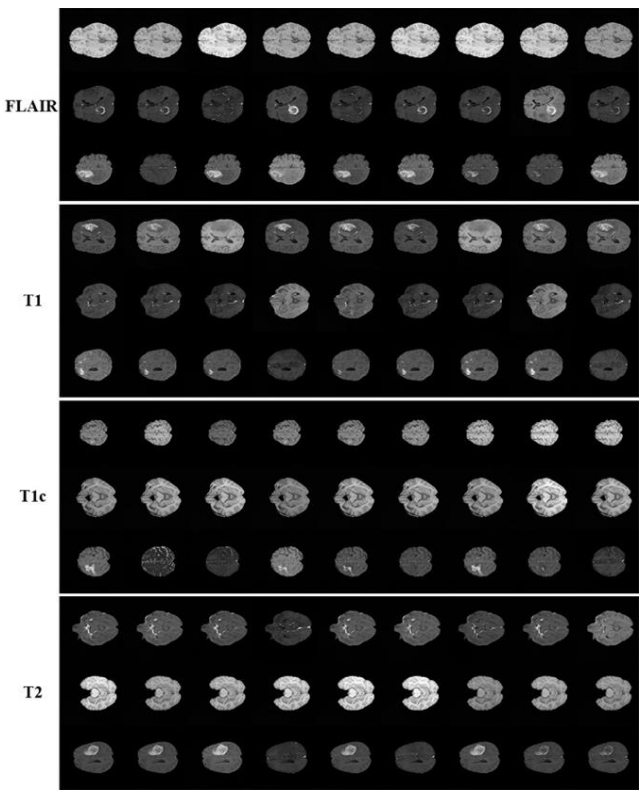

Fig. 9 StarGAN V2 training process generating medical images

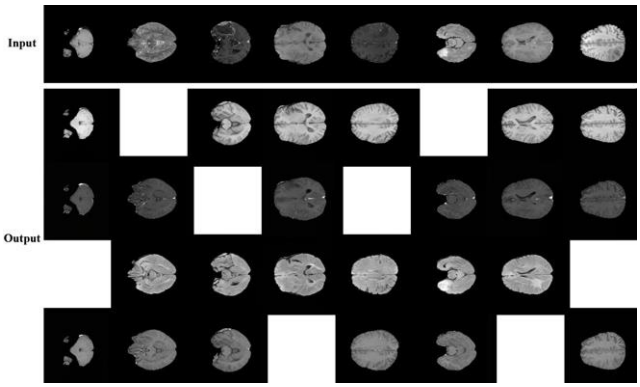

Fig. 10 Here we show the results of the medical imaging. The first row shows the input data, and the remaining data represents the data generated buy the model.

erence image. In Fig. 10, we show the results of generating medical images after model training.

We evaluate both the visual quality and the diversity of generated images using Frechét inception distance (FID) [42] and learned perceptual image patch similarity (LPIPS) [43]. We compute FID and LPIPS for every pair of image domains within a dataset and report their average values. In the section of validation of medical image segmentation experiments, We evaluated the segmentation effect by calculating Dice Similarity Coefficient (Dice).

Multimodal unsupervised image translation is implemented in MUNIT, so in this paper, we try to use this as a base to train different models, but the training effect is not satisfactory. We can summarize the reasons for the unsatisfactory training results of MUNIT here. Since the focus of MUNIT is still to do image translation from a domain to another one, which is similar to CycleGAN, so if we want to do multimodal translation, for example, the dataset used in this paper has four modalities, and we want to translate them to each other, then we have to train 12 generators to complete it. The style part in MUNIT is only to learn the details of the target image. We give a further explanation, for example, training a T1-T2 translator in MUNIT (T1 and T2 are MRI of the same patient in different weighted states), when the role of the style part is to input a T2 image and let the generator learn the original T1 image according to the style of the target T2 image, If the input T2 image is a medical image with tumor, then the T2 image generated based on T1 image is a medical image with tumor, but this is not meaningful for the follow-up task we want to do. On the other hand, the model trained in this paper is a multimodal model, if a represents the data of T1 modality, then B is the data of the remaining modal medical images mixed together, so it is not effective to use MUNIT to achieve multimodal data augmentation. In this paper, to validate this idea and to train with CycleGAN under the same conditions, we show it in Fig. 11 Since CycleGAN is not focused on implementing multi-domain translation, CycleGAN is not used as a baseline in this paper and is only shown as a reference.

Here, we provide further explanation of the metrics

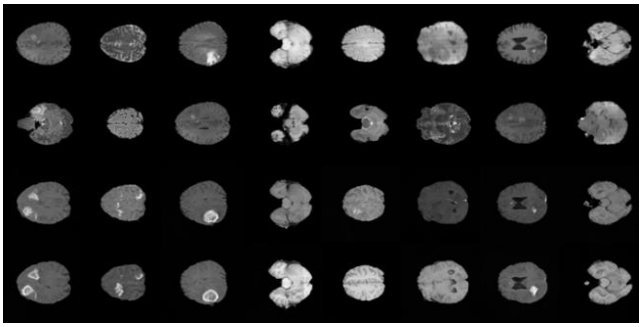

Fig. 11 CycleGAN training BRATS process to generate images

Table 4 Latent-guided synthesis

Method	iteration = 50000		iteration = 100000	
	FID	LPIPS	FID	LPIPS
MUNIT	73.896	0.059	83.772	0.075
StarGAN v2	29.281	0.163	23.953	0.119
ours	**27.990**	**0.172**	**21.551**	**0.159**

used in this paper, Frechet Inception Distance score (FID) is a method to calculates the distance between the real image and the feature vectors of the generated image.

$$FID(x, g) = \|\mu_x - \mu_g\|^2 + Tr\left(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x \Sigma_g}\right) \quad (21)$$

where μ and Σ indicate the mean and variance of the clusters. r and g subscripts represent the clustering points from the real dataset and the generated dataset, respectively. *Tr* represents the tracking operator. μ_x, Σ_x are the mean and covariance matrices of the 2048-dimensional feature vector set of the real image collection at the output of Inception Net-V3 [44], respectively, and μ_g, Σ_g are the mean and covariance matrices of the 2048-dimensional feature vectors of the image collection generated at the output of Inception Net-V3, separately. A lower FID value implies a closer match between the generated distribution and the real image distribution, and if the real images used for testing are high in clarity and variety, sit also implies a high quality and good diversity of the generated images. We record the value of the FID in Table 4 and Table 6.

Learned Perceptual Image Patch Similarity (LPIPS), also known as "perceptual loss", is used to measure the difference between two images, which is a metric that learns to generate a reverse mapping of the image to Ground Truth. A lower value of LPIPS means that the two images are more similar. Given a Ground Truth image reference block x and a noise-containing image distortion block x_0, the perceptual similarity measure is formulated as follows.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw} - \hat{y}^l_{0hw} \right) \right\|_2^2 \quad (22)$$

where d is the distance x_0 from x. The feature stack is extracted from the L layer and unit-normalized in the channel dimension. The vectors are used to $w_l \in R^{c_l}$ deflate

Table 5 Single mode latent-guide synthesis

Contrast	FID			
	FLAIR	T1	T1c	T2
FLAIR	——	23.012	26.100	22.023
T1	21.219	——	20.054	20.307
T1c	20.081	20.750	——	20.321
T2	25.124	22.990	23.281	——

Table 6 Reference-guided synthesis

Method	iteration = 50000		iteration = 100000	
	FID	LPIPS	FID	LPIPS
MUNIT	99.361	0.037	186.965	0.088
StarGAN v2	24.727	0.125	36.510	0.105
ours	**24.534**	**0.151**	**20.540**	**0.116**

Table 7 Single mode reference-guided synthesis

Contrast	FID			
	FLAIR	T1	T1c	T2
FLAIR	——	21.587	22.609	20.962
T1	20.573	——	23.132	21.510
T1c	22.278	22.661	——	21.365
T2	20.681	21.466	21.229	——

Table 8 Results of segmentation data

Model	Original dataset			Ours		
	DiceWT	DiceTC	DiceET	DiceWT	DiceTC	DiceET
$model_1$	0.86	0.74	0.76	0.90	0.80	0.84
$model_2$	0.89	0.86	0.82	0.93	0.89	0.88

the number of activated channels, and eventually the L2 distance is calculated. Finally, it is averaged over the space and summed over the channels.

Dice was used to evaluate segmentation results. The Dice measures the overlap between the ground truth and the automatic segmentation which is described by

$$Dice(P,T) = \frac{|P \wedge T|}{\frac{1}{2}(|P| + |T|)} \tag{23}$$

where P and T are the pixel-prediction area of the brain tumor and the ground truth of the brain tumor, respectively. The signal '$\wedge$' represents the intersection of two areas above, and the signal '+' means taking the union of two areas. The value of Dice is between 0 and 1. The larger the value, the better the quality of the segmentation. We show this in Table 8.

In Tables 4 and 6, we show the total scores of LPIPS and FID for latent-guided synthesis and reference-guided synthesis, respectively, and Tables 5 and 7 show the FID scores for medical images with other modalities generated from a single modality. FID denotes that the distance between the two distributions of the real and generated images

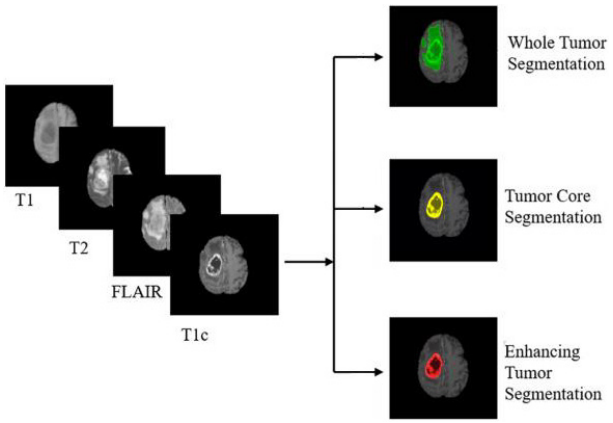

Fig. 12 Segmentation area of BRATS.

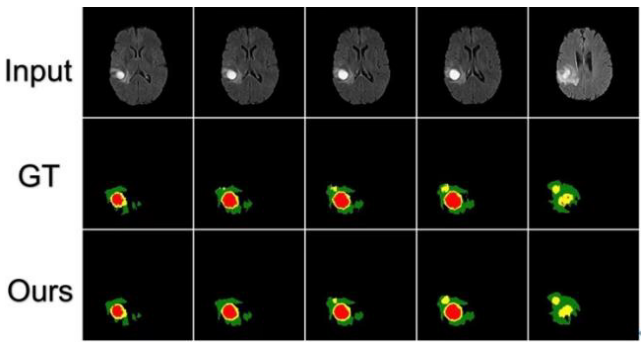

Fig. 13 The segmentation results after data augmentation, compared to the manually delineated ground truth. Input represents the medical image after data augmentation. ours represents the output of our data after the segmentation model.

(lower is better), in contrast to LPIPS which measures the diversity of the generated images (higher is better). The data from our experiments can verify that the idea of medical image data augmentation based on GAN proposed in this paper is desirable.

In Fig. 12, we give the regions to be segmented, Whole Tumor segmentation (WT), Tumor Core segmentation (TC), and Enhancing Tumor segmentation (ET) in the Dice metric, whose higher values indicate better segmentation ability. Finally in Fig. 13, we show some segmentation results of the demonstration case compared to the manually delineated ground truth (GT).

It can be visualized from Fig. 13 that there is almost no difference between the training segmentation results and the GT images. Table 8 shows the experimental results of $model_1$ and $model_2$ compared with the original data, indicating that the medical image data proposed in this paper have positive effects after augmentation and are of research value.

5. Conclusion

In this paper, an improved method based on the principle of StarGAN V2 is designed and applied to medical images.

A small amount of original images can be used to augment the dataset and the augmented multimodal medical images can be applied to the segmentation task to improve the segmentation effect. The experiments in this paper verify that the StarGAN V2-based network designed in this paper has high performance, and also demonstrate that the multimodal medical images obtained after data augmentation training are effective for unimodal medical image segmentation. The training model in this paper achieves high segmentation results and effectively improves the sample dependence of deep learning. In addition, this paper can be used to augment other image datasets and help solve the problem of insufficient image classification and target detection datasets. Follow-up work can focus on optimizing the network structure and the relevant characteristics of each mode to achieve higher segmentation accuracy of brain tumors in medical images using small sample training.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant 62266004.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, ands A.A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Proc Cvpr IEEE, pp.5967–5976, 2017. (DOI: 10.1109/Cvpr.2017.632).

[2] W.H. Xia, Y.Yang, and J.-H. Xue, "Unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement," Neural Networks, vol.131, pp.50–63, 2020. (DOI: 10.1016/j.neunet.2020.07.023).

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," Proc Cvpr IEEE, pp.105–114, 2017. (DOI: 10.1109/Cvpr.2017.19).

[4] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to Image Translation for Domain Adaptation," 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), pp.4500–4509, 2018. (DOI: 10.1109/Cvpr.2018.00473).

[5] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised Diverse Colorization via Generative Adversarial Networks," Lect Notes Artif Int, vol.10534, pp.151–166, 2017. (DOI: 10.1007/978-3-319-71249-9_10).

[6] G. Mariani, et al., "BAGAN: Data augmentation with balancing GAN," arXiv preprint arXiv:1803.09655, 2018.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," Comm. ACM, vol.63, no.11, pp.139–144, 2020. (DOI: 10.1145/3422622).

[8] M. Mirza, et al., "Conditional generative adversarial nets," arXiv preprint arXiv: 1411.1784, 2014.

[9] J.-Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," IEEE I Conf Comp Vis., pp.2242–2251, 2017. (DOI: 10.1109/Iccv.2017.244).

[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Cvpr), 2020. arXiv preprint arXiv: 1912.01865.

[11] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-Shot Unsupervised Image-to-Image Translation," 2019 IEEE/Cvf International Conference on Computer Vision (Iccv 2019), pp.10550–10559, 2019. (DOI: 10.1109/Iccv.2019.01065).

[12] C. Qi, et al., "SAG-GAN: Semi-supervised attention-guided GANs for data augmentation on medical images," arXiv preprint arXiv:2011.07534, 2020.

[13] H. Montenegro, W. Silva, and J.S. Cardoso, "Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis," IEEE Access, vol.9, pp.148037–148047, 2021. (DOI: 10.1109/ACCESS.2021.3124844).

[14] G. Ramachandra, et al., "GAN augmentation: Augmenting training data using generative adversarial networks," arXiv preprint arXiv: 1810.10863, 2017.

[15] Q. Li, Z. Gao, Q. Wang, J. Xia, H. Zhang, H. Zhang, H. Liu, and S. Li, "Glioma segmentation with a unified algorithm in multimodal MRI images," IEEE Access, vol.6, pp.9543–9553, 2018. (DOI: 10.1109/ACCESS.2018.2807698).

[16] M.I. Sharif, J.P. Li, M.A. Khan, and M.A. Saleem, "Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images," Pattern Recogn Lett, vol.129, pp.181–189, 2020. (DOI: 10.1016/j.patrec.2019.11.019).

[17] S. Alqazzaz, X. Sun, X. Yang, and L. Nokes, "Automated brain tumor segmentation on multimodal MR image using SegNet," Computational Visual Media, vol.5, pp.209–219, 2019. (DOI: 10.1007/s41095-019-0139-y).

[18] E.G. Van Meir, C.G. Hadjipanayis, A.D. Norden, H.-K. Shu, P.Y. Wen, and J.J. Olson, "Exciting new advances in neuroncology: the avenue to a cure for malignant glioma," CA: a cancer journal for clinicians, vol.60, no.3, pp.166–193, 2010. (DOI: 10.3322/caac.20069).

[19] K.L. Tseng, et al., "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," arXiv preprint arXiv:1704.07754, 2017.

[20] Q. Li, Z. Yu, Y. Wang, and H. Zheng, "TumorGAN: A Multi-Modal Data Augmentation Framework for Brain Tumor Segmentation," Sensors, vol.20, no.15, 4203, 2020. (DOI: 10.3390/s20154203).

[21] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A Phase Congruency and Local Laplacian Energy Based Multi-Modality Medical Image Fusion Method in NSCT Domain," IEEE Access, vol.7, pp.20811–20824, 2019. (DOI: 10.1109/ACCESS.2019.2898111).

[22] K. Wang, M. Zheng, H. Wei, G. Qi, and Y. Li, "Multi-Modality Medical Image Fusion Using Convolutional Neural Network and Contrast Pyramid," Sensors, vol.20, no.8, 2169, 2020. (DOI: 10.3390/s20082169).

[23] Z.Q. Zhu, H. Wei, G. Hu, Y. Li, G. Qi, and N. Mazur, "A Novel Fast Single Image Dehazing Algorithm Based on Artificial Multiexposure Image Fusion," IEEE Trans. Instrum. Meas., vol.70, 99, 2020. (DOI: 10.1109/Tim.2020.3024335).

[24] M. Zheng, G. Qi, Z. Zhu, Y. Li, H. Wei, and Y. Liu, "Image Dehazing by An Artificial Image Fusion Method based on Adaptive Structure Decomposition," IEEE Sensors J., vol.20, no,14, pp.8062–8072, 2020. (DOI: 10.1109/JSEN.2020.2981719).

[25] C. Kaushal and A. Singla, "Automated segmentation technique with self-driven post-processing for histopathological breast cancer images," CAAI Transactions on Intelligence Technology, 2020. (DOI: 10.1049/trit.2019.0077).

[26] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," Caai Transactions on Intelligence Technology, vol.6, no.1, pp.25–45, 2021. (DOI: 10.1049/cit2.12028).

[27] H. Zhang, V. Sindagi, and V.M. Patel, "Image De-Raining Using a Conditional Generative Adversarial Network," IEEE Trans. Circuits Syst. Video Technol., vol.30, no.11, pp.3943–3956, 2020. (DOI: 10.1109/TCSVT.2019.2920407).

[28] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," Proc Cvpr IEEE, pp.8789–8797, 2018. (DOI: 10.1109/Cvpr.2018.00916).

[29] M. Arjovsky, et al., "Wasserstein GAN," arXiv preprint arXiv:1701.07875, 2017.
[30] I. Gulrajani, et al., "Improved training of Wasserstein GANs," arXiv preprint arXiv:1704.00028, 2017.
[31] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," 2017 IEEE International Conference on Computer Vision (Iccv), pp.1510–1519, 2017 (DOI: 10.1109/Iccv.2017.167).
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (Cvpr), pp.770–778, 2016 (DOI: 10.1109/CVPR.2016.90).
[33] V. Nair, et al., "Rectified linear units improve restricted Boltzmann machines," ICML, https://icml.cc/Conferences/2010/papers/432.pdf, 2010.
[34] L. Mescheder, et al., "Which training methods for gans do actually converge?" arXiv preprint arXiv:1801.04406v2, 2018.
[35] A.L. Maas, et al., "Rectifier nonlinearities improve neural network acoustic models," Proc. ICML, vol.30, 1, 2013 (DOI: 10.1.1.693.1422).
[36] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation," Lect Notes Comput Sc, vol.11207, pp.179–196, 2018. (DOI: 10.1007/978-3-030-01219-9_11).
[37] Q. Wang, et al., "ECA-net: Efficient channel attention for deep convolutional neural networks," arXiv preprint arXiv:1910.03151, 2020.
[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), pp.7132–7141, 2018. (DOI: 10.1109/Cvpr.2018.00745).
[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention, vol.9351, pp.234–241, 2015. (DOI: 10.1007/978-3-319-24574-4_28).
[40] A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Brainles 2018, vol.11384, pp.311–320, 2019. (DOI: 10.1007/978-3-030-11726-9_28).
[41] S. Bakas, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," arXiv preprint arXiv:1811.02629 (02018), 2018.
[42] M. Hensel, et al., "GANs trained by a two time-scale update rule converge to a local nash equilibrium," arXiv preprint arXiv:1706.08500, 2017.
[43] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr), pp.586–595, 2018. (DOI: 10.1109/Cvpr.2018.00068).
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr), pp.2818–2826, 2016. (DOI: 10.1109/Cvpr.2016.308).

Yue Peng received her B.S. degree in 2016 and is currently a master's student at Guangxi University. Her research interests in the current research include image processing, machine learning.

Zuqiang Meng graduated from Central South University with a Ph.D. in Computer Application Technology in 2004 and was qualified as a postdoctoral fellow at the Institute of Computing, Chinese Academy of Sciences in 2009. He is currently working as a professor at Guangxi University, and his main research interests are cross-modal intelligence, granular computing and data mining.

Lina Yang received her M.S. degree in Computer Science from University of Malaya in 2011, Ph.D. degree in Software Engineering from University of Macau in 2015, and postdoctoral research work in Institute of Science and Technology, University of Macau in 2016. She is currently working as a professor at Guangxi University, her main research interests include Cross-modal Intelligence, granular computing and data mining.