Anatomic and Molecular MR Image Synthesis Using Confidence Guided CNNs

Pengfei Guo, *Student Member, IEEE,* Puyang Wang, *Student Member, IEEE,* Rajeev Yasarla, *Student Member, IEEE,* Jinyuan Zhou, Vishal M. Patel, *Senior Member, IEEE,* and Shanshan Jiang, *Member, IEEE*

***Abstract*— Data-driven automatic approaches have demonstrated their great potential in resolving various clinical diagnostic dilemmas in neuro-oncology, especially with the help of standard anatomic and advanced molecular MR images. However, data quantity and quality remain a key determinant, and a significant limit of the potential applications. In our previous work, we explored the synthesis of anatomic and molecular MR image networks (SAMR) in patients with post-treatment malignant gliomas. In this work, we extend this through a confidence-guided SAMR (CG-SAMR) that synthesizes data from lesion contour information to multi-modal MR images, including T1-weighted (T_1w), gadolinium enhanced T_1w (Gd-T_1w), T2-weighted (T_2w), and fluid-attenuated inversion recovery (FLAIR), as well as the molecular amide proton transfer-weighted (APTw) sequence. We introduce a module that guides the synthesis based on a confidence measure of the intermediate results. Furthermore, we extend the proposed architecture to allow training using unpaired data. Extensive experiments on real clinical data demonstrate that the proposed model can perform better than current the state-of-the-art synthesis methods. Our code is available at https://github.com/guopengf/CG-SAMR.**

***Index Terms*— Generative adversarial network, Confidence guidance, Multi-modal MR image synthesis, Glioma, Segmentation.**

I. Introduction

Glioblastoma (GBM) is the most malignant and frequently occurring type of primary brain tumor in adults. Despite the development of various aggressive treatments, patients with GBM inevitably suffer tumor recurrence with an extremely poor prognosis [1]–[4]. The dilemma in the clinical management of post-treatment patients remains precise assessment of the treatment responsiveness. Magnetic resonance imaging (MRI) is considered the best non-invasive assessment method of GBM treatment responsiveness [5]–[7]. Compared to anatomic MRI, such as T1-weighted (T_1w), gadolinium enhanced T_1w (Gd-T_1w), T2-weighted (T_2w), and fluid-attenuated inversion recovery ($FLAIR$) images, amide proton transfer-weighted (APTw) MRI is a novel molecular imaging technique that is able to generate endogenous contrast to detect mobile proteins and peptides in vivo. APTw MRI has been proven by many researchers to positively influence clinical management [8]–[20]. Its application in patients with brain tumors was approved by the FDA recently. With the help of convolutional neural networks (CNNs), data-driven medical image analysis methods have provided exciting solutions in the neuro-oncologic community [21], [22]. Several studies have demonstrated that CNNs-based methods outperform humans on fine-grained classification tasks but require a large amount of accurately annotated, diversity-rich data [23], [24]. It is usually impractical to collect large MRI datasets, especially for advanced MR image data. Furthermore, obtaining aligned lesion annotations on the corresponding co-registered multi-modal MR images (namely, paired training data) is costly, since expert radiologists are required to label the data, monitor the image preprocessing, and verify the annotation based on extensive professional knowledge. While deploying conventional data augmentations, such as rotation, flipping, random cropping, and distortion, during training partly mitigates these issues, the performance of CNNs models is still limited because the diversity of datasets is compromised [25].

Many novel approaches, including generative adversarial networks (GAN), have been explored for generating more realistic data. Goodfellow et al. [26] proposed GAN and first applied it to synthesize photo-realistic images. Isola et al. [27] and Wang et al. [28] further investigated conditional GAN and achieved an impressive solution to image-to-image translation problems. Several generative models have been successfully proposed for MRI synthesis. Dar et al. [29] developed conditional GANs for spatially registered/unregistered multi-contrast MR images. Nguyen et al. [30] and Chartsias et al. [31] proposed CNNs-based architectures to synthesize cross-modality MR images. Cordier et al. [32] further used a generative model for multi-modal MR images of brain tumors from a single label map. However, their inputs were conventional MRI modalities, and the diversity of the synthesized images was limited by the training images. Moreover, the method is not yet capable of producing manipulated outputs. Shin et al. [33] adopted Pix2Pix [27] to transfer brain anatomy and lesion segmentation maps to multi-modal MR images with brain tumors. This approach is capable of producing manipulated outputs and realistic brain anatomy for multiple MRI sequences, but it does not consider significant differences in radiographic features between anatomic and molecular MRI.

This work was supported in part by grants from the National Institutes of Health (R01CA228188) and the National Science Foundation (1910141).

Pengfei Guo, Puyang Wang, Rajeev Yasarla, and Vishal M. Patel are with the Whiting School of Engineering, Johns Hopkins University, (e-mail: {pguo4, pwang47, ryasarl1, vpatel36}@jhu.edu).

Jinyuan Zhou and Shanshan Jiang are with the School of Medicine, Johns Hopkins University, (e-mail: {jzhou2,sjiang21}@jhmi.edu).

Moreover, lesions with diverse and complicated patterns may need extra supervision during synthesis [34], [35]. In this scenario, more new methods need to be explored for multi-modal MR images and optimized for lesion regions.

In our previous work, the synthesis of anatomic and molecular MR images networks (SAMR) [36], a novel generative model was proposed to simultaneously synthesize a diverse set of anatomic T_1w, Gd-T_1w, T_2w, and $FLAIR$ images, as well as molecular APTw images. SAMR [36] is a GAN-based approach. It takes arbitrarily manipulated lesion masks facilitated by brain atlases generated from training data as the input and consists of a stretch-out up-sampling module, a segmentation consistency module, and multi-scale label-wise discriminators. In this paper, we extend SAMR [36] by incorporating extra supervision on the latent features and corresponding confidence information to further improve the synthetic performance, especially in lesion regions. Intuitively, directly providing the estimated synthesized images (i.e. intermediate results) to the subsequent layers of the networks may propagate errors to the final synthesized images. With the confidence map module, the proposed algorithm is capable of measuring an uncertainty metric of the intermediate results and blocking the flow of incorrect estimation. To this end, we formulate a joint task of estimating the confidence score at each pixel location of the intermediate results and synthesizing realistic multi-modal MR images, namely confidence guided SAMR (CG-SAMR). Figure 1(a) presents an overview of the proposed method corresponding to training using paired data. Furthermore, to overcome the insufficiency of paired training data, we modified the network to allow unsupervised training, namely unpaired CG-SAMR (UCG-SAMR). In other words, the proposed approach does not require aligned pairs of lesion segmentation maps and multi-modal MR images during training. This is achieved by adding an extra GAN which reverses the synthesis process to a segmentation task. The schematic of the proposed method that corresponds to training using unpaired data is presented in Figure 1(b). In summary, this paper makes the following contributions:

- A novel GAN-based model, called CG-SAMR, is proposed to synthesize high quality multi-modal anatomic and molecular MR images with controllable lesion information.
- Confidence scores of each sequence measured during synthesis are used to guide the subsequent layers for better synthesis performance.
- To increase the diversity of the synthesized data, rather than explicitly using white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) masks, we leverage the atlas of each sequence to provide brain anatomic information in CG-SAMR.
- We demonstrate the feasibility of extending the application of the CG-SAMR network to unpaired data training.
- Comparisons are made against several recent state-of-the-art paired/unpaired synthesis approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by various components of the proposed method.

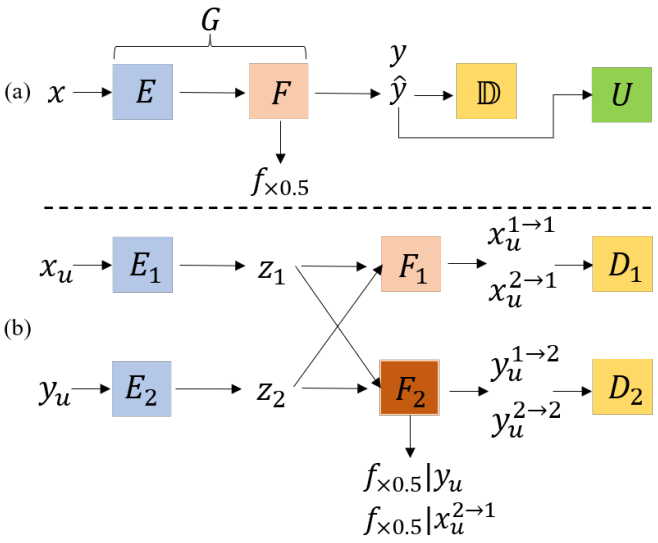

Fig. 1. Overview of the proposed frameworks. (a) The schematic of the proposed method that corresponds to training using paired data. $\boldsymbol{G}$ is the generator of a GAN which consists of an encoder, E and a decoder, F. $\mathbb{D}$ represents multi-scale label-wise discriminators. $\boldsymbol{U}$ is the U-net lesion segmentation module. $\boldsymbol{f_{\times 0.5}}$ represents feature maps at a scale of $\times 0.5$. (b) The schematic of the proposed method that corresponds to training using unpaired data. $\boldsymbol{E_1}$, and $\boldsymbol{E_2}$ are two encoders that maps input to the latent codes. $\boldsymbol{F_1}$ is a decoder that maps the latent codes to lesion segmentation maps and anatomic prior (domain 1). $\boldsymbol{F_2}$ is a decoder that has the same network architecture as the decoder of CG-SAMR. $\boldsymbol{F_1}$ can generate two types of instances: (1) instances from the reconstruction stream $\boldsymbol{x_u^{1\to 1} = F_1(E_1(x_u))}$, and (2) instances from the cross-domain stream $\boldsymbol{x_u^{2\to 1} = F_1(E_2(y_u))}$. $\boldsymbol{F_2}$ maps the latent codes to multi-modal MR images (domain 2). Similarly, $\boldsymbol{F_2}$ also can generate two types of instances. We denote the feature maps used for confidence estimation in $\boldsymbol{F_2}$ as $\boldsymbol{f_{\times 0.5}|y_u}$ when decoding the latent code obtained by encoding $\boldsymbol{y_u}$. $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ are two discriminators for domain 1 and domain 2, respectively.

The rest of the paper is organized as follows. Section II provides a review of some related works. Details of the proposed method are given in Section III. Implementation details, experimental results, and the ablation study are given in Section IV. Finally, Sections V and VI conclude the paper with a discussion and summary.

II. Related Works

The goal of MR image synthesis is to generate target images with realistic radiographic features [37]. Technically, MR image synthesis can be achieved by a generative model that translates the source domain to the MR image domain. The source domain usually belongs to noise or different modalities/contrast types (e.g., from CT images to MR images or from T_1w images to T_2w images). In what follows, we review some recent studies on this topic and applications of modeling uncertainty in CNNs.

A. Conventional Methods

Conventional medical image synthesis methods include intensity-based methods and registration-based methods [38]. Intensity-based methods essentially learn a transformation function that maps source intensities to target intensities. Roy

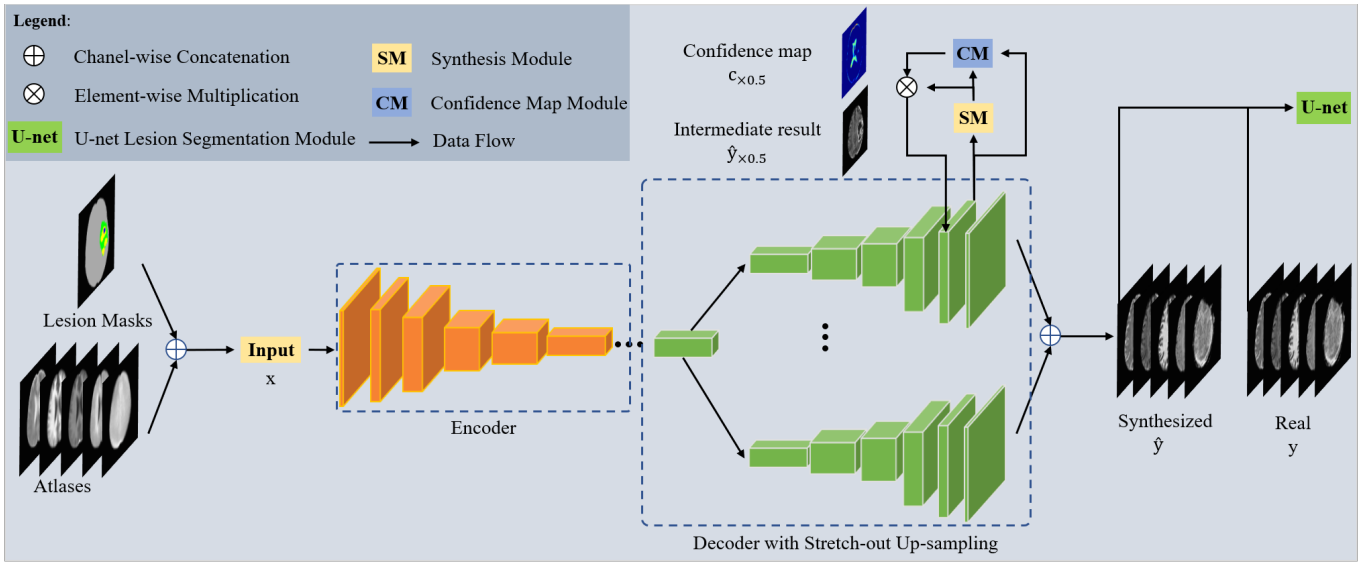

Fig. 2. An overview of the proposed CG-SAMR network. The goal of the CG-SAMR network is to produce realistic multi-modal MR images given the corresponding lesion masks and atlases. The orange blocks indicate the encoder part. The green blocks represent the decoder part with stretch-out up-sampling, in which we perform a customized synthesis for each MRI sequence. The synthesis module produces the intermediate results for each branch of stretch-out up-sampling and is denoted as SM. CM represents the confidence map module that computes confidence maps to guide the subsequent networks. The U-net lesion segmentation module regularized the encoder-decoder part to produce lesion regions with correct radiographic features.

et al. [39] proposed an example-based approach that relied on sparse reconstruction from image patches to achieve contrast synthesis and further extended it under the setting of patch-based compressed sensing [40]. Joy *et al.* [41] leveraged random forest regression to learn the nonlinear intensity mappings for synthesizing full-head T_2w images and $FLAIR$ images. Huang *et al.* [42] proposed a geometry-regularized joint dictionary learning framework to synthesize cross-modality MR images. For registration-based methods, the synthesized images are generated by the registration between source images and target co-registered images [43]. Cardoso *et al.* [44] further extended this idea to synthesize expected intensities in an unseen image modality by a template-based, multi-modal, generative mixture-model.

B. CNNs-based Methods

With the development of deep learning, CNNs-based medical image synthesis methods have shown significant improvements over the conventional methods of image synthesis. Rather than using patch-based methods [45], [46], Sevetlidis *et al.* [47] introduced a whole image synthesis approach that relied on a CNNs-based autoencoder architecture. Nguyen *et al.* [30] and Chartsias *et al.* [31] proposed CNNs-based architectures that integrated intensity features from images to synthesize cross-modality MR images. Joyce et al. [29] presented a multi-input encoder-decoder neural network model that leveraged learned modality invariant latent embedding to perform MR image synthesis in both single and multi-input settings and demonstrated the robustness of dealing with both missing and misaligned data. Dar et al. [48] addressed the data scarcity problem in training CNNs models for accelerated MRI by a transfer-learning approach. They leveraged the pre-trained networks on a large natural images dataset and fine-tuned on a small amount of MR images to achieve domain transfer between natural and MR images. Various GAN-based methods have also been used for medical image analysis [49], [50]. Dar et al. [51] proposed mustGAN, which is a multi-stream approach that integrates information across multiple source images via a mixture of multiple one-to-one streams and a joint many-to-one stream for multi-modal MR image synthesis. Shin et al. [33] adopted pix2pix [27] to transfer brain anatomy and lesion segmentation maps to multi-modal MR images with brain tumors, which showed the benefit of using brain anatomy prior, such as white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) masks, in MR image synthesis.

One major challenge of image synthesis is that paired source/target images are required during training which are expensive to acquire. Recent advances in GAN-based architectures, cycle-consistent adversarial networks (CycleGAN) [52] and unsupervised image-to-image translation networks (UNIT) [53] have outlined to a promising direction for cross-modality biomedical image synthesis using unpaired source/target images. Wolterink *et al.* [54] leveraged cycle consistency to achieve bidirectional MR/CT image synthesis. A conditional GAN-based approach has been proposed by Dar et al. [55] to tackle the multi-contrast MR image synthesis in the scenarios of registered and unregistered images by introducing the cycle consistency loss. Chartsias *et al.* [56] proposed a two stage framework for MR/CT image synthesis and demonstrated that the synthesized data can further improve the segmentation performance. Zhang *et al.* [57] and Huo *et al.* [38] introduced SynSeg-Net to achieve bidirectional synthesis and anatomy segmentation. In their approach, the source domain was the MR images and segmentation labels, while the target domain is CT images. Inspired by these works,

we also added an extra GAN-based network to CG-SAMR and leveraged cycle consistency to allow the training using unpaired data.

C. Modeling Uncertainty in CNNs

Many recent approaches have modeled the uncertainty in CNNs and used it to benefit the network in different applications. Kendall *et al.* [58] leveraged the Bayesian deep learning models to demonstrate the benefit of modeling uncertainty on semantic segmentation and depth regression tasks. In [59], Kendall *et al.* extended the previous work [58] to multi-task learning by proposing a multi-task loss function that maximized the Gaussian likelihood with homoscedastic uncertainty. Yasarla *et al.* [60] and Jose *et al.* [61] modeled the aleatoric uncertainty as maximum likelihood inference on image restoration and ultrasound image segmentation tasks, respectively. Inspired by these works, we introduce a novel loss function to measure the confidence score of the intermediate synthesis results and guide the subsequent networks of CG-SAMR by the estimated confidence scores.

III. Methodology

Figure 2 gives an overview of the proposed encoder and decoder parts of CG-SAMR framework. By incorporating multi-scale label-wise discriminators and shape consistency-based optimization, the generator aims to produce meaningful, high-quality anatomical and molecular MR images with diverse and controllable lesion information. While applying 3D convolution operations might reflect the reality of data, the output of the proposed method is multi-modal MRI image slices, since the voxel size between anatomical and molecular MRI in the axial direction is significantly different and re-sampling to isotropic resolution can severely degrade the image quality. Detailed imaging parameters are given in Section IV-A. In what follows, we describe key parts of the network and training processes using paired and unpaired data.

A. Multi-modal MRI Generation

Our generator architecture is inspired by the models proposed by Johnson et al. [62] and Wang et al. [28]. The generator network, consists of two components (see Figure 2): an encoder and a decoder with a stretch-out up-sampling module. Let the set of multi-modal MR images be denoted as $\mathcal{Y}$ and the corresponding set of lesion segmentation maps and anatomic prior as $\mathcal{X}$. In CG-SAMR, the anatomic prior corresponds to multi-modal atlas. Details of atlas generation are provided in Section IV-A and Supplementary Figure 2 shows an example of generated atlases. The generator aims to synthesize multi-modal MR images $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$. Unlike many deep learning-based methods that directly synthesize MR images from input, we first estimate the intermediate synthesis results $\hat{y}_{\times 0.5}$ (0.5 scale size of y) and the corresponding confidence map $c_{\times 0.5}$, then use them to guide the synthesis of the final output $\hat{y}$. The input x is passed through the encoder module to obtain the latent feature maps. Then, the same latent feature maps are passed through each branch of the stretch-out up-sampling block to perform customized synthesis.

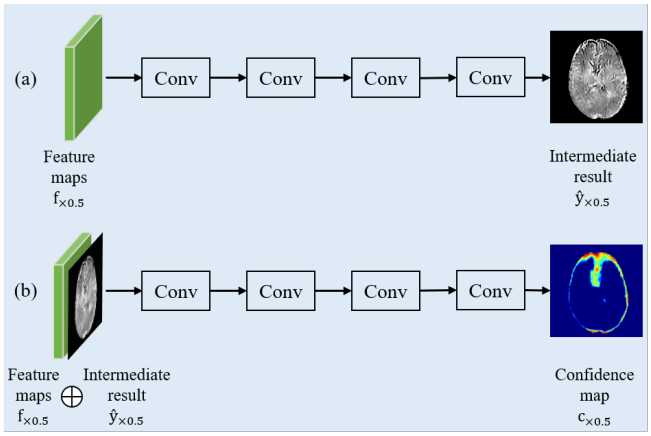

Fig. 3. (a) Synthesis module. (b) Confidence map module. Conv represents a convolution block that contains a convolutional layer, a batch normalization layer, and a Rectified Linear Units (ReLU) activation. $\oplus$ is the channel-wise concatenation. The visualization of confidence maps are generated by $(\mathbf{1} - \boldsymbol{c}_{\times 0.5})$ for better illustration of the uncertain region.

The encoder part (orange blocks in Figure 2) consists of a fully-convolutional module with five layers and subsequent three residual learning blocks (ResBlock) [63]. We set the kernel size and stride equal to 7 and 1, respectively, for the first layer. For the purpose of down-sampling, instead of using maximum-pooling, the stride of other four layers was set to 2. Rectified Linear Unit (ReLu) activation and batch normalization were sequentially added after each layer. To learn better transformation functions and representations through a deeper perception, the depth of the encoder network is increased by three ResBlocks [25], [63]. We observed significantly different radiographic features between anatomic and molecular MR images, which vastly increased the difficulty of simultaneous synthesis. To address this issue, the decoder part (green blocks in Figure 2) consists of three ResBlocks and a stretch-out up-sampling module that contains five same sub-modules designed to utilize the latent representations from the preceding ResBlock and perform customized synthesis for each MR sequence. Each sub-module contains a symmetric architecture with a fully-convolutional module in the encoder. All convolutional layers are replaced by transposed convolutional layers for up-sampling. The synthesized multi-modal MR images are produced from each sub-model. Details of convolutional layers in the generator are shown in Supplementary Methods and Supplementary Table 1.

B. Synthesis and Confidence Map Modules

The synthesis networks are prone to generating incorrect radiographic features at or near the edges, since they usually involve diverse and complicated patterns. Thus, special attention to those regions where the network tends to be uncertain can improve the MR image synthesis task. To address this issue, a synthesis module and a confidence map module are added to each branch of the stretch-out up-sampling block

(see Synthesis Module (SM) and Confidence Map Module (CM) in Figure 2). Specifically, we estimate the intermediate synthesis results at a 0.5 scale size of the final output by SM and measured the confidence map which gives attention to the uncertain regions by CM. The confidence score at each pixel is a measurement of certainty about the intermediate results computed at each pixel. On confidence maps, regions where the network was certain about the synthesized intensity showed high confidence values (i.e. close to 1), while the network assigned low confidence scores (i.e. close to 0) for those pixels where it was uncertain. To this end, we can block the erroneous regions by combing confidence maps and the intermediate results. Thus, the masked intermediate results are returned to the subsequent networks, which makes the network more attentive in the uncertain regions.

As shown in Figure 3, feature maps at a scale of $\times 0.5$ ($f_{\times 0.5}$) are given as input to SM to compute the intermediate results of each MR sequence at a scale of $\times 0.5$. SM is a sequence of four convolutional blocks. Then, we feed the estimated intermediate results and the feature maps as inputs to CM to compute the confidence scores at every pixel. CM is also a sequence of four convolutional blocks. Details of the convolutional layers in SM and CM are given in Supplementary Table 2. Finally, the confidence-masked intermediate results (i.e. the element-wise multiplication between $\hat{y}_{\times 0.5}$ and $c_{\times 0.5}$) combined with feature maps at a scale of $\times 0.5$ are fed back to the network to guide the subsequent layers to produce the final output. Inspired by modeling the data dependent aleatoric uncertainty [58], [59], we define the confidence map loss as follows

$$\begin{aligned} \mathcal{L}_{\text{CM}}(f_{\times 0.5}) &= c_{\times 0.5} \otimes \|\hat{y}_{\times 0.5} - y_{\times 0.5}\|_1 - \lambda_{\text{cm}} C, \\ \hat{y}_{\times 0.5} &= \text{SM}(f_{\times 0.5}), \\ c_{\times 0.5} &= \text{CM}(f_{\times 0.5} \oplus \hat{y}_{\times 0.5}), \\ C &= \sum_i \sum_j \log(c^{ij}_{\times 0.5}), \end{aligned} \tag{1}$$

where $\otimes$, $\oplus$ are the element-wise multiplication and the channel-wise concatenation, respectively. $c^{ij}_{\times 0.5}$ represents the confidence score at the ith row, jth column of the confidence map $c_{\times 0.5}$. $\hat{y}_{\times 0.5}$ represents the intermediate synthesis results produced by the decoder part. In $\mathcal{L}_{\text{CM}}$, the first term minimizes the L1 difference between $\hat{y}_{\times 0.5}$ and $y_{\times 0.5}$, and the values of $c_{\times 0.5}$ as well. To avoid a trivial solution (i.e. $c^{ij}_{\times 0.5} = 0, \forall i, j$), we introduced the second term as a regularizer. λ_{cm} is a constant adjusting the weight of this regularization term C. A similar loss was used for image restoration and ultrasound segmentation tasks in [61], [64]. To the best of our knowledge, our method is the first attempt to introduce this kind of loss in MR synthesis tasks.

C. Multi-scale Label-wise Discriminators

In order to achieve a large receptive field in discriminators without introducing deeper networks, we adopt multi-scale PatchGAN discriminators [27], which have identical network architectures but accept multi-scale inputs [28]. To distinguish between real and synthesized images, conventional discriminators operate on the whole input. However, optimizing the generator to produce realistic images in each regions of interest (ROI) cannot be guaranteed by discriminating on holistic images, since the difficulty of synthesizing images in different regions varies. To address this issue, we introduce label-wise discriminators. Based on the radiographic features, original lesion segmentation masks were reorganized into three ROIs, including background, normal brain, and lesion. As shown in Figure 4, the input of each discriminator is masked by its corresponding ROI. Since the proposed discriminators are in a multi-scale setting, for each ROI there are two discriminators that operate on the original and a down-sampled $\times 0.5$ scale. Thus, there are in total six discriminators for three ROIs and we refer to these set of discriminators as $\mathbb{D} = \{D_1, D_2, D_3, D_4, D_5, D_6\}$. In particular, $\{D_1, D_2\}$,$\{D_3, D_4\}$, and $\{D_5, D_6\}$ operate on the original and down-sampled versions of background, normal brain, and lesion, respectively. An overview of the proposed discriminators is given in Figure 4. The objective function corresponding to the discriminators $\mathcal{L}_{\text{GAN}}(G, D_k)$ is as follows

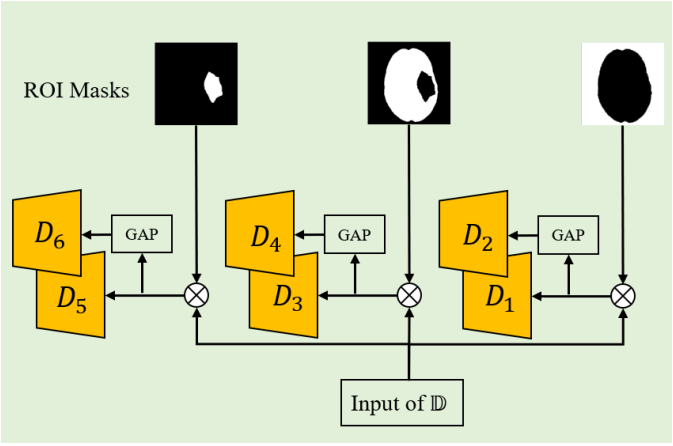

Fig. 4. The overview of multi-scale label-wise discriminators. ROI masks are produced from reorganized input lesion masks. We denote $\otimes$ as the element-wise multiplication operation. GAP is the global average pooling that generates a 0.5 scale size of input. $\mathbb{D}$ is a set of discriminators.

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(G, D_k) =& \mathbb{E}_{(x', y')}[\log D_k(x', y')] \\ &+ \mathbb{E}_{x'}[\log(1 - D_k(x', G'(x)))], \\ \mathcal{L}_{\text{GAN}}(G, D) =& \sum_{k=1}^{6} \mathcal{L}_{\text{GAN}}(G, D_k), \end{aligned} \tag{2}$$

where x and y are paired input and real multi-modal MR images, respectively. $x' \triangleq m_k \otimes x$, $y' \triangleq m_k \otimes y$, and $G'(x) \triangleq m_k \otimes G(x)$, where $\otimes$ denotes element-wise multiplication and m_k corresponds to the ROI mask. For simplicity, we omit the down-sampling operation in this equation.

D. Training Using Paired Data

A multi-task loss is designed to train the generator and the discriminators in an adversarial setting. Rather than only using the conventional adversarial loss $\mathcal{L}_{\text{GAN}}$, we also adopt a feature matching loss $\mathcal{L}_{\text{FM}}$ [28] to stabilize training, which optimizes the generator to match these intermediate representations from

the real and the synthesized images in multiple layers of the discriminators. For discriminators, $\mathcal{L}_{\text{FM}}(G, D_k)$ is defined as follows

$$\begin{aligned}\mathcal{L}_{\text{FM}}(G, D_k) &= \sum_{i}^{T} \frac{1}{N_i} \left[\| D_k^{(i)}(x', y') - D_k^{(i)}(x', G'(x) \|_2^2 \right] \\ \mathcal{L}_{\text{FM}}(G, D) &= \sum_{k=1}^{6} \mathcal{L}_{\text{FM}}(G, D_k),\end{aligned} \tag{3}$$

where $D_k^{(i)}$ denotes the ith layer of the discriminator D_k, T is the total number of layers in D_k and N_i is the number of elements in the ith layer. When we perform lesion segmentation on images, it is worth noting that there is a consistent relation between the prediction and the real image serving as input for the generator. In most of the lesions, the annotations of different labels are usually irregularly interlaced, which causes ambiguity for synthesizing realistic MR images. To tackle this problem, we propose a lesion shape consistency loss $\mathcal{L}_{\text{SC}}$ by adding a U-net [65] segmentation module (see Figure 2) that regularizes the generator to obey this consistency relation. We adopt a Generalized Dice Loss (GDL) [66] to measure the difference between the predicted and real segmentation maps which is defined as follows

$$\text{GDL}(R, S) = 1 - \frac{2 \sum_i^N r_i s_i}{\sum_i^N r_i + \sum_i^N s_i}, \tag{4}$$

where R denotes the ground truth and S is the segmentation result. r_i and s_i represent the ground truth and predicted probability maps at each pixel i, respectively. N is the total number of pixels. The lesion shape consistency loss $\mathcal{L}_{\text{SC}}$ is then defined as follows

$$\mathcal{L}_{\text{SC}}(U) = \text{GDL}(s, U(y)) + \text{GDL}(s, U(G(x))), \tag{5}$$

where $U(y)$ and $U(G(x))$ represent the predicted lesion segmentation probability maps by taking y and $G(x)$ as inputs in the segmentation module, respectively. s denotes the ground truth lesion segmentation map. The final multi-task objective function for training CG-SAMR is defined as

$$\begin{aligned}\min_{\text{G,U}}(\max_{\text{D}} \mathcal{L}_{\text{GAN}}(G, D)) &+ \lambda_1 \mathcal{L}_{\text{FM}}(G, D) \\ &+ \lambda_2 \mathcal{L}_{\text{SC}}(U) + \lambda_3 \mathcal{L}_{\text{CM}}(f_{\times 0.5}),\end{aligned} \tag{6}$$

where λ_1, λ_2 and λ_3 are the three parameters that control the importance of each loss.

E. Training Using Unpaired Data

Figure 5 shows the schematic of the proposed method that corresponds to training using unpaired data. Our framework is based on the proposed CG-SAMR network and additional GANs: $\text{GAN}_1 = \{E_1, F_1, D_1\}$ and $\text{GAN}_2 = \{E_2, F_2, D_2\}$. Denote the set of lesion segmentation maps and anatomic prior as domain 1 and the set of multi-modal MR images as domain 2. Here, we denote **unpaired** instances in domain 1 and 2 as x_u and y_u, respectively. In GAN_1, D_1 aims to evaluate whether the translated unpaired instances are realistic. It outputs true for real instances sampled from the domain 1 and false for instances generated by F_1. As shown in Figure 5,

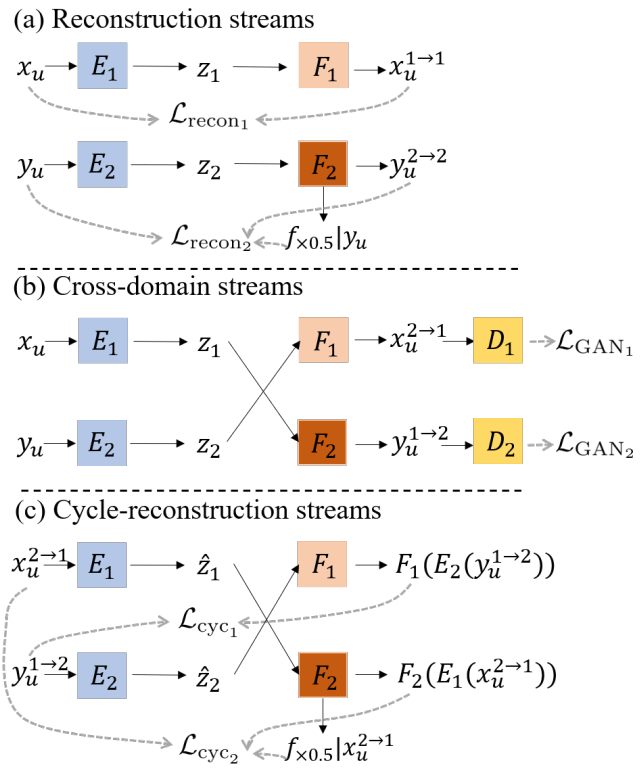

Fig. 5. The overview of the proposed method in (a) reconstruction streams, (b) cross-domain streams, and (c) cycle-reconstruction streams under the training using unpaired data.

F_1 can generate two types of instances: (1) instances from the reconstruction stream $x_u^{1\to1} = F_1(E_1(x_u))$, and (2) instances from the cross-domain stream $x_u^{2\to1} = F_1(E_2(y_u))$. We have similar properties in GAN_2, but the decoder F_2 is replaced by the corresponding decoder part in CG-SAMR. Thus, we can realize confidence-guided customized synthesis for each MR sequence under unpaired data training. The objective functions for reconstruction streams (Figure 5a) are defined as follows

$$\begin{aligned}\mathcal{L}_{\text{recon}_1} &= \|x_u - x_u^{1\to1}\|_1, \\ \mathcal{L}_{\text{recon}_2} &= \|y_u - y_u^{2\to2}\|_1 + \mathcal{L}_{\text{CM}}(f_{\times0.5}|y_u),\end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{CM}}$ is defined in equation (1) and F_2 is a decoder network with the same architecture as used in CG-SAMR. We denote the feature maps used for $\mathcal{L}_{\text{CM}}$ in F_2 as $f_{\times0.5}|y_u$ when decoding the latent code obtained by encoding y_u. The objective functions of cross-domain streams (Figure 5b) can be expressed as follows

$$\begin{aligned}\mathcal{L}_{\text{GAN}_1} &= \mathbb{E}_{(x_u)}[\log D(x_u)] + \mathbb{E}_{(z_2)}[\log(1 - D_1(x_u^{2\to1})], \\ \mathcal{L}_{\text{GAN}_2} &= \mathbb{E}_{(y_u)}[\log D(y_u)] + \mathbb{E}_{(z_1)}[\log(1 - D_2(y_u^{1\to2})],\end{aligned} \tag{8}$$

where z_1 and z_2 are the latent codes, $z_1 = E_1(x_u)$, $z_2 = E_2(y_u)$. Relying on the reconstruction stream and adversarial training (i.e. cross-domain streams) cannot guarantee learning of the desired mapping function. To reduce the number of possible mapping functions, we require the learned mapping functions to obey cycle-consistent constraints (i.e. $x_u \to y_u^{1\to2} \to F_1(E_2(y_u^{1\to2})) \approx x_u$) [52]. The objective functions for cycle-reconstruction streams (Figure 5c) are defined as

TABLE I

QUANTITATIVE COMPARISON. THE QUALITY OF THE SYNTHESIZED DATA UNDER PAIRED DATA TRAINING IS MEASURED BY PIXEL ACCURACY FOR LESION REGIONS, INCLUDING EDEMA, CAVITY, AND TUMOR. THE UNIT IS IN PERCENT (%). A SYNTHESIZED PIXEL WAS COUNTED CORRECT IF THE ABSOLUTE DIFFERENCE WAS WITHIN 16 OF THE GROUND TRUTH INTENSITY VALUE. THE QUALITY OF HOLISTIC SYNTHESIZED IMAGES WAS MEASURED BY SSIM AND PSNR.

	Pix2Pix [27]					Pix2PixHD [28]					Shin et al. [33]					SAMR [36]					CG-SAMR (ours)				
	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR
	Edema	Cavity	Tumor			Edema	Cavity	Tumor			Edema	Cavity	Tumor			Edema	Cavity	Tumor			Edema	Cavity	Tumor		
APTw	50.9	43.0	48.2	0.766	17.8	54.9	42.5	51.6	0.765	18.2	45.3	40.4	41.8	0.770	17.2	66.1	53.4	63.3	0.782	18.1	67.3	51.6	64.4	0.782	18.4
T_1w	54.5	56.3	49.5	0.789	17.5	53.8	52.8	47.9	0.794	17.5	73.2	72.2	68.4	0.900	25.0	73.3	69.4	67.7	0.821	19.4	76.4	67.8	71.3	0.834	20.2
$FLAIR$	52.2	41.8	45.3	0.778	20.4	47.4	37.3	46.8	0.774	20.6	59.7	42.8	51.9	0.810	22.1	75.6	62.0	68.2	0.798	22.5	78.5	67.8	71.7	0.809	23.5
T_2w	52.9	52.7	43.4	0.799	20.2	51.4	59.4	47.3	0.794	20.3	65.7	56.2	56.5	0.845	22.8	77.1	78.0	71.3	0.822	22.6	81.5	77.9	74.6	0.836	23.5
Gd-T_1w	70.5	57.7	38.8	0.747	20.8	72.3	58.3	37.9	0.754	21.0	74.9	65.3	39.4	0.818	22.8	81.7	67.8	64.7	0.790	22.7	83.5	69.5	63.1	0.801	23.4
Avg.	56.2	50.3	45.0	0.776	19.3	56.0	50.1	46.3	0.776	19.5	63.8	55.4	51.6	**0.829**	**22.0**	74.8	66.1	67.0	0.803	21.1	**77.4**	**66.9**	**69.0**	0.812	21.8

TABLE II

QUANTITATIVE RESULTS CORRESPONDING TO IMAGE SEGMENTATION WHEN THE SYNTHESIZED DATA IS USED FOR DATA AUGMENTATION. FOR EACH EXPERIMENT, THE FIRST ROW REPORTS THE PERCENTAGE OF SYNTHESIZED/REAL DATA FOR TRAINING AND THE NUMBER OF INSTANCES OF SYNTHESIZED/REAL DATA IN PARENTHESES. EXP.3 REPORTS THE RESULTS OF THE BASELINE TRAINED ONLY BY REAL DATA.

Exp.1: 50% Synthesized + 50% Real (1080 + 1080)			
	Dice Score		
	Edema	Cavity	Tumor
Pix2Pix [27]	0.594	0.453	0564
Pix2PixHD [28]	0.597	0.532	0.566
Shin et al. [33]	0.734	0.700	0.733
SAMR [36]	0.789	0.819	0.813
CG-SAMR (ours)	**0.807**	**0.840**	**0.835**
Exp.2: 25% Synthesized + 75% Real (540 + 1080)			
Pix2Pix [27]	0.600	0.505	0.565
Pix2PixHD [28]	0.642	0.510	0.662
Shin et al. [33]	0.675	0.647	0.708
SAMR [36]	0.750	0.779	0.773
CG-SAMR (ours)	**0.760**	**0.793**	**0.772**
Exp.3: 0% Synthesized + 100% Real (0 + 1080)			
Baseline	**0.647**	**0.610**	**0.672**

follows

$$\begin{aligned}\mathcal{L}_{\text{cyc}_1} =& \|x_u - F_1(E_2(y_u^{1\to 2}))\|_1,\\ \mathcal{L}_{\text{cyc}_2} =& \|y_u - F_2(E_1(x_u^{2\to 1}))\|_1 + \mathcal{L}_{\text{CM}}(f_{\times 0.5}|x_u^{2\to 1}).\end{aligned} \tag{9}$$

The overall objective function used to train the UCG-SAMR in the setting of unpaired data training is defined as follows

$$\begin{aligned}G^* =& \min_{\{E_1,F_1,E_2,F_2\}} \max_{\{D_1,D_2\}} \mathcal{L}_{\text{domain}_1} + \mathcal{L}_{\text{domain}_2}, \text{ where}\\ \mathcal{L}_{\text{domain}_1} =& \mathcal{L}_{\text{recon}_1} + \mathcal{L}_{\text{GAN}_1} + \mathcal{L}_{\text{cyc}_1}, \text{ and}\\ \mathcal{L}_{\text{domain}_2} =& \mathcal{L}_{\text{recon}_2} + \mathcal{L}_{\text{GAN}_2} + \mathcal{L}_{\text{cyc}_2}.\end{aligned} \tag{10}$$

IV. EXPERIMENTS AND RESULTS

In this section, we first discuss the data acquisition and training details. Then, the experimental setup, evaluations of the proposed synthesis methods against a set of recent state-of-the-art approaches, and comprehensive ablation studies are presented.

A. Data Acquisition

This retrospective study was approved by the Institutional Review Board (IRB) and conducted in accordance with the U.S. Common Rule, and the need for a consent form was waived. Patient inclusion criteria were: at least 20 years old; initial diagnosis of pathologically proven primary malignant glioma; status post initial surgery and chemoradiation. Data from 100 patients were re-analyzed in this study [13], [36], [67]. MRI scans were acquired on a 3T human MRI scanner (Achieva; Philips Medical Systems) by using a body coil excite and a 32-channel phased-array coil for reception [67]. T_1w, Gd-T_1w, T_2w, $FLAIR$, and APTw MRI sequences were collected for each patient. Image parameters for APTw can be summarized as: field of view (FOV), 212 $\times$ 212 $\times$ 66 mm^3; resolution, 0.82 $\times$ 0.82 $\times$ 4.4 mm^3; and size of matrix, 256 $\times$ 256 $\times$ 15. Other anatomic MRI sequences were acquired with image parameters: FOV, 212 $\times$ 212 $\times$ 165 mm^3; resolution, 0.41 $\times$ 0.41 $\times$ 1.1 mm^3; and size of matrix, 512 $\times$ 512 $\times$ 150. Co-registration between APTw and anatomic sequences [68], skull stripping [69], N4-bias field correction [70], and MRI standardization [71] were performed sequentially. The schematic of the data preprocessing pipeline is shown in Supplementary Figure 3. After preprocessing, the final volume size of each sequence was 256 $\times$ 256 $\times$ 15. For every collected volume, lesions were manually annotated by an expert neuroradiologist into three labels: edema; cavity; and tumor. A lesion of postoperative malignant glioma was first annotated to cover the region of abnormal intensities on $FLAIR$ MR images. Then, the radiologist further labeled "cavity" (including surgical cavity and cavity with liquefactive necrosis), and "tumor" (including active and inactive tumor residual) within the lesion. Then, the remainder of the lesion was defined as "edema". Notably, the prior/follow-up MR images and daily progress notes were reviewed with caution, in order to make sure the dynamic changes within the residual tumor and the prior/ongoing therapies were comprehensively considered. Then, a multivariate template construction tool [72] was used to create the group average of each sequence (atlas) from volumes used for training. Fifteen hundred instances from 100 patients with a size of 256 $\times$ 256 $\times$ 5 were extracted from volumetric data, where 5 corresponds to five MRI sequences. For every instance, the corresponding atlas slice and two adjunct atlas slices in the axial direction were extracted to provide the prior of human brain anatomy in paired data training. The WM, GM, CSF probability masks were also extracted to provide anatomic prior used in the scenario of unpaired data training by SPM12 [73]. We split these instances into 1080 (72 patients) for training, 150 (10 patients) for validation and 270 (18 patients) for testing. The

patient-based data was split to ensure that training, validation and testing data did not include the instances from the same patient.

B. Implementation Details

The hyperparameter selection in deep neural networks, especially for GAN, is computationally intensive [74]. We performed commonly used one-fold validation in deep learning research on the validation dataset to select the optimal combination of hyperparameters [55], [75], [76]. The CG-SAMR synthesis model was trained based on the final objective function equation (6) using the Adam optimizer [72]. λ_1, λ_2 and λ_3 were set equal to 5, 1 and 1, respectively. Hyperparameters were set as follows: constant learning rate of 2×10^{-4} for the first 250 epochs then linearly decaying to 0; 500 maximum epochs; batch size of 8. λ_{cm} in equation (1) initially was set equal to 0.1. When the mean of scores in confidence maps $c_{\times 0.5}$ was greater than 0.7, λ_{cm} was set equal to 0.03. Hyperparameters for unpaired data training were set as follows: constant learning rate of 2×10^{-4} for the first 400 epochs then linearly decaying to 0; 800 maximum epochs; batch size of 1. To further evaluate the effectiveness of the synthesized MRI sequences on data augmentation, we leveraged U-net [65] to train lesion segmentation models. U-net [65] was trained by the Adam optimizer [72]. Hyperparameters were set as follows: constant learning rate of 2×10^{-4} for the first 100 epochs then linearly decaying to 0; 200 maximum epochs; batch size of 16. In the segmentation training, all the synthesized data were produced from randomly manipulated lesion masks by CG-SAMR. For comparison methods, training procedures and hyperparameters were adopted from their original publications.

C. Results of Training Using Paired Data

We evaluated the performance of our method against the following recent state-of-the-art generic synthesis methods: Pix2Pix [27], Pix2PixHD [28] as well as MRI synthesis methods: Shin et al. [33], and SAMR [36]. We used pixel accuracy [27], [28], [52] to compare the performance in lesion regions. In particular, we calculate the difference between the synthesized data and the corresponding ground truth data. A pixel mapping was counted correct if the absolute difference was within 16 of the ground truth intensity value. The structural similarity index measure (SSIM) and peak-signal-to-noise ratio (PSNR) were also introduced to the quality evaluation of holistic synthesized images. Table I shows the quantitative performance of different methods in terms of pixel accuracy, SSIM, and PSNR. As can be seen from this table, our method clearly outperformed the present state-of-the-art synthesis algorithms at lesion regions. Figure 6 presents the qualitative comparisons of the synthesized multi-modal MRI sequences and zoomed-in images around the gadolinium enhanced region from different methods. It can be observed that Pix2Pix [27] and Pix2PixHD [28] were less optimal for the synthesis of realistic looking human brain MR images. In Figure 6 (b)(c), the disparities included abnormal brain anatomic structures (i.e., an emerging cerebral ventricle dilation, disproportionately atrophic corpus callosum) and unreasonable radiographic features in the lesion regions. Moreover, the frontal lobes showed misleading hypointensity (deep blue), while normal appearing normal brain parenchyma always presented as green or very light blue (the APTw intensity is around 0%). Shin et al. [33] produced realistic brain anatomic structures for anatomic MRI sequences. However, there is an obvious disparity between the synthesized

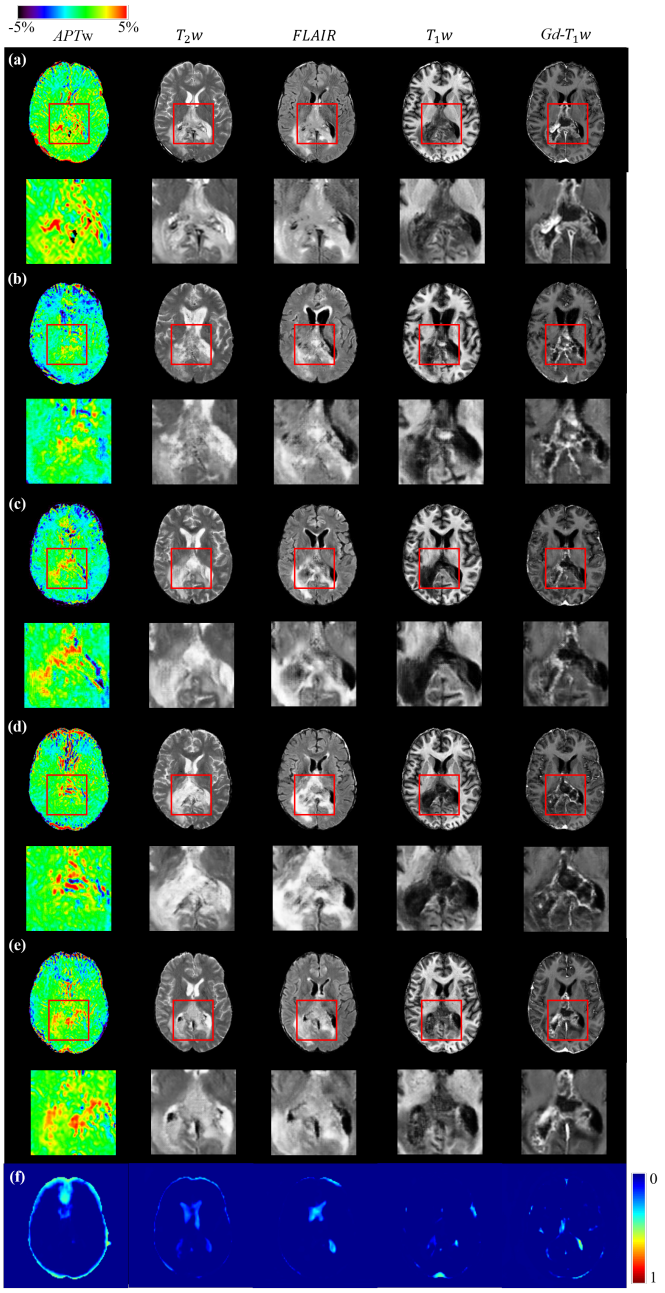

Fig. 6. Qualitative comparison of different methods under paired data training. The same lesion mask is used to synthesize images from different methods. (a) Real data (ground truth). (b) Pix2Pix [27]. (c) Pix2PixHD [28]. (d) Shin et al. [33]. (e) CG-SAMR (ours). (f) Confidence maps from CG-SAMR. The visualization of confidence maps are generated by $(\mathbf{1} - \boldsymbol{c}_{\times \mathbf{0.5}})$ for better illustration of the uncertain region. The second row of each approach shows zoomed-in images of the gadolinium-enhanced regions of the images in the first row, which are indicated by red boxes.

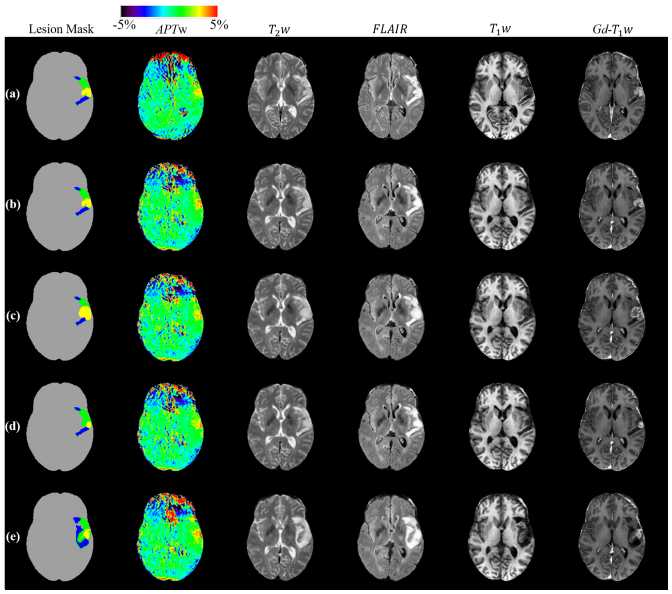

Fig. 7. Examples of lesion mask manipulations in CG-SAMR. (a) Real images (ground truth). (b) Synthesized images from the original mask. (c) Synthesized images when tumor size is increased by 100%. (d) Synthesized images when tumor size is reduced to 50%. (e) Synthesized images when the lesion mask is replaced by another one. In lesion masks, gray, green, yellow, and blue represent normal brain, edema, tumor, and cavity, respectively.

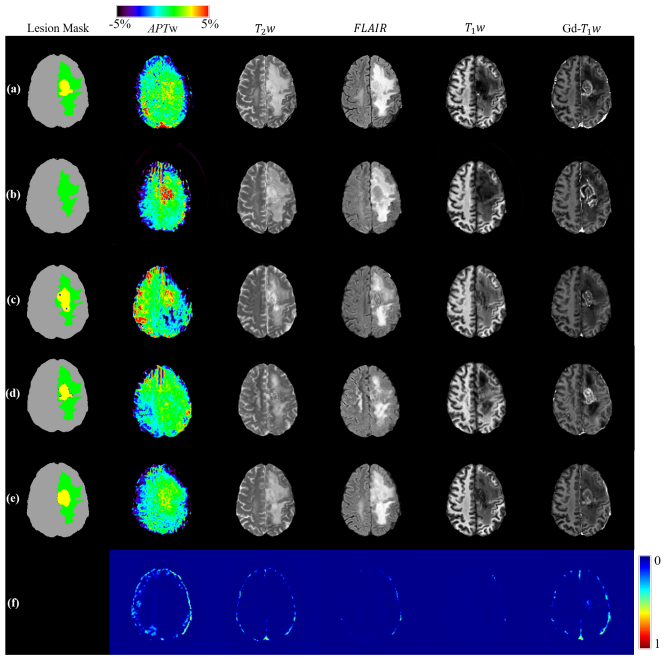

Fig. 8. Qualitative comparison of segmentation and synthesis performance under unpaired data training. (a) Real data (ground truth). (b) CycleGAN [52]. (c) UNIT [28]. (d) SynSeg-Net [38]. (e) UCG-SAMR (ours). (f) Confidence maps from UCG-SAMR. The visualization of confidence maps are generated by $(\mathbf{1} - \boldsymbol{c}_{\times \mathbf{0.5}})$ for better illustration of the uncertain region. In lesion masks, gray, green, yellow, and blue represent normal brain, edema, tumor, and cavity, respectively.

and real APTw images, especially in the lesion regions (see red boxes in Figure 6 (d)). Our proposed method produced more reasonable radiographic features of lesions and a more realistic anatomic structure. We provide the visualization of image differences between synthesized images and the ground truth in Supplementary Figure 4 and an additional qualitative comparison in Supplementary Figure 5.

To further evaluate the quality of the synthesized MR images, we performed data augmentation by using the synthesized images in training and then performed lesion segmentation. The dice score was used as an evaluation metric to measure the performance of different methods. The data augmentation by synthesis was evaluated by the improvement in lesion segmentation models. We arbitrarily controlled lesion information to synthesize different amounts of data for augmentation. Supplementary Figure 1 shows the flowchart of lesion mask manipulation. To simulate the practical use of data augmentation, we conducted experiments by using utilizing all real data. In each experiment, we varied the percentage of the synthesized data to observe the contribution to data augmentation. Table II shows the calculated segmentation performance. Compared to the baseline experiment that only used real data, the synthesized data from pix2pix [27] and pix2pixHD [28] degraded the segmentation performance. The performance was improved when the synthesized data from Shin et al. [33] and SAMR [36] were used for segmentation, but the proposed method outperforms the other methods by a large margin. Figure 7 demonstrates the robustness of the proposed model under different lesion mask manipulations (e.g. changing the size of tumor and even reassembling lesion information between lesion masks). As can be seen from this figure, our method is robust to various lesion mask manipulations.

D. Results of Training Using Unpaired Data

We called the proposed method under unpaired data training as UCG-SAMR and evaluated its performance against the recent state-of-the-art generic synthesis methods (CycleGAN [52] and UNIT [53]) as well as another MRI synthesis methods (SynSeg-Net [38]). Table III shows the quantitative synthesis performance of different methods in term of pixel accuracy, SSIM, and PSNR. As can be seen from this table, our method outperformed the other state-of-the-art synthesis algorithms. UCG-SAMR improves pixel accuracy by 17.7% 7.6%, and 4.3% at lesion regions (average value of pixel accuracies from edema, cavity and tumor) compared to CycleGAN [52], SynSeg-Net [38] and UNIT [53], respectively. Figure 8 presents an example to show the qualitative comparison of the segmentation and synthesis performance using unpaired data. Compared to CycleGAN [52], SynSeg-Net [38] and UNIT [53], UCG-SAMR yields synthesized images with remarkable visual similarity to reference image and error was obviously suppressed in lesion regions as shown in the visualization of image differences between the synthesized images and the ground truth under unpaired data training (see Supplementary Figure 6). Table IV shows the comparison of segmentation performance for different methods. Due to the introduction of an extra segmentation network, we can observe that SynSeg-Net [38] exhibited superior capacity and reached the performance upper bound (i.e. supervised training by real paired data in Table II Exp.3). Facilitated by high-quality synthesis, the

TABLE III

QUANTITATIVE COMPARISON. THE QUALITY OF THE SYNTHESIZED DATA UNDER UNPAIRED DATA TRAINING IS MEASURED BY PIXEL ACCURACY FOR LESION REGIONS, INCLUDING EDEMA, CAVITY, AND TUMOR. THE UNIT IS IN PERCENT (%). A SYNTHESIZED PIXEL WAS COUNTED CORRECT IF THE ABSOLUTE DIFFERENCE WAS WITHIN 16 OF THE GROUND TRUTH INTENSITY VALUE. THE QUALITY OF HOLISTIC SYNTHESIZED IMAGES WAS MEASURED BY SSIM AND PSNR.

	CycleGAN [52]					SynSeg-Net [38]					UNIT [53]					UCG-SAMR (ours)				
	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR	Pixel Accuracy			SSIM	PSNR
	Edema	Cavity	Tumor			Edema	Cavity	Tumor			Edema	Cavity	Tumor			Edema	Cavity	Tumor		
APTw	51.5	32.8	39.8	0.636	15.8	40.5	38.0	43.9	0.638	15.4	41.9	33.3	50.3	0.628	13.5	51.6	37.3	44.8	0.643	15.6
T_1w	35.2	23.5	34.2	0.739	19.1	54.7	54.6	45.2	0.789	20.1	67.7	73.0	63.0	0.801	19.8	66.4	60.8	68.9	0.852	22.7
$FLAIR$	57.7	33.0	35.8	0.696	19.7	49.0	37.7	44.4	0.707	19.8	65.0	38.9	59.1	0.714	20.1	66.2	37.2	58.1	0.743	20.7
T_2w	67.6	5.7	55.1	0.729	19.5	63.8	50.1	56.7	0.755	20.2	64.4	12.8	60.8	0.735	20.4	69.3	48.7	60.0	0.781	21.2
Gd-T_1w	47.9	45.1	22.9	0.703	19.8	69.0	59.1	32.1	0.734	20.4	79.1	40.3	39.3	0.733	20.4	72.7	66.3	45.8	0.768	21.2
Avg.	52.0	28.0	37.6	0.701	18.8	55.4	47.9	44.5	0.726	19.2	63.6	39.7	54.5	0.722	18.8	**65.2**	**50.1**	**55.5**	**0.757**	**20.3**

TABLE IV

QUANTITATIVE EVALUATION OF THE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS UNDER UNPAIRED DATA TRAINING.

	Dice Score		
	Edema	Cavity	Tumor
CycleGAN [52]	0.233	0.138	0.120
UNIT [53]	0.533	0.368	0.509
SynSeg-Net [38]	**0.630**	**0.607**	**0.591**
UCG-SAMR (our)	0.553	0.366	0.527

TABLE V

ABLATION STUDY OF DESIGNED MODULES IN DATA AUGMENTATION BY SYNTHESIS IS ACCESSED BY DICE SCORE. ABLATION STUDY OF DESIGNED MODULES IN TERM OF SYNTHESIS QUALITY IS ACCESSED BY PIXEL ACCURACY. THE REPORTED VALUE IS PIXEL ACCURACY IN THE LESION REGION (I.E. THE UNION OF EDEMA, CAVITY, AND TUMOR) AS PERCENT (%). A SYNTHESIZED PIXEL WAS COUNTED CORRECT IF THE ABSOLUTE DIFFERENCE WAS WITHIN 16 OF THE GROUND TRUTH INTENSITY VALUE.

	Dice Score			Pixel Accuracy					
	Edema	Cavity	Tumor	APTw	T_1w	$FLAIR$	T_2w	Gd-T_1w	Avg.
w/o Stretch-out	0.684	0.695	0.678	62.7	66.3	66.6	70.8	73.1	67.9
w/o Multi-label D	0.758	0.798	0.787	63.6	74.3	73.5	74.6	77.8	72.8
w/o Atlas	0689	0.710	0.705	61.7	66.4	69.6	73.7	74.1	69.1
w/o $\mathcal{L}_{SC}$	0.733	0.794	0.772	63.6	72.0	71.6	76.1	76.5	72.4
w/o $\mathcal{L}_{CM}$	0.789	0.819	0.813	64.0	73.0	73.3	77.5	78.5	73.3
CG-SAMR	**0.807**	**0.840**	**0.835**	**64.4**	**75.4**	**76.6**	**81.0**	**79.6**	**75.4**

proposed UCG-SAMR network achieved the second-best performance against other models, as can be seen from Table IV.

E. Ablation Study

We conducted a comprehensive ablation study to separately evaluate the effectiveness of using a stretch-out up-sampling module in the decoder network, label-wise discriminators, the atlas, lesion shape consistency loss $\mathcal{L}_{SC}$, and confidence map loss $\mathcal{L}_{CM}$ in the proposed method. We evaluated each designed module based on two aspects: (1) the effectiveness of data augmentation by the synthesized data, and (2) the contribution to the synthesis quality. For the former, we used the same experimental setting as exp.1 in Table II. The effectiveness of modules for data augmentation by synthesis is reported in Table V. The pixel accuracy in Table V shows the contribution of designed modules in the MR image synthesis of different sequences. We could observe that the two experiments showed a similar trend. Losing the customized reconstruction for each sequence (stretch-out up-sampling module) can severely degrade the synthesis quality. We found that when the atlas was not used in our method, it significantly affected the synthesis quality due to the lack of human brain anatomy prior. Moreover, dropping either $\mathcal{L}_{SC}$ or label-wise discriminators in the training also reduced the performance, since the shape consistency loss and the specific supervision on ROIs were not used to optimize the generator to produce more realistic images. In addition, dropping the confidence loss $\mathcal{L}_{CM}$ can lead to performance degradation, since the supervision on the intermediate results and attention of uncertain regions during synthesis can provide improved results. The comparison between Supplementary Figure 4(e) and (f) demonstrates that CG-SAMR exhibits better ability of suppressing error on uncertain regions.

V. DISCUSSION

The potential to leverage lesion masks and anatomic structures prior by deep learning-based models, and thus, to synthesize multi-modal anatomic and molecular MR images was investigated in this study of the post-treatment malignant gliomas. We demonstrated that the proposed method CG-SAMR incorporated with a confidence map loss $\mathcal{L}_{CM}$ can effectively improve synthesis quality in lesion regions and benefit data augmentation by synthesized data. While many previous methods require paired multi-modal MR images for training, our UCG-SAMR method utilizes cycle-consistency loss for training to synthesize MR images from unpaired data. Extensive evaluations were performed for two distinct scenarios where training images were paired and unpaired. The experiments showed that the proposed method highlights a promising direction for the synthesis of multi-modal MR images with rich radiographic features for post-treatment malignant gliomas and facilitates future studies about data-driven methods for human patients.

On the basis of our experimental results, our proposed approach has the following unique advantages. First, while many studies that have developed GAN synthesis methods have been published, we focused on the specific problem of synthesizing multi-modal MR images from lesion masks. On the foundation of our previous method SAMR [36], we leveraged an atlas of each sequence to provide brain anatomy prior rather than subject-specific WM, GM, and CSF masks (used in many previous works). This promotes a wide range of diversity for the synthesized data. The stretch-out up-sampling module performs customized synthesis for each MR sequence. Label-wise discriminators are designed to provide specific supervision on each ROI. The lesion shape consistency loss is proposed to regularize the generator to produce realistic

lesions. Second, we introduced a confidence-guided training strategy. A synthesis module and a confidence map module were added on each branch of the stretch-out up-sampling block. Specifically, we estimated the intermediate synthesis results by SM and measured the confidence map which provides attention to the uncertain regions by CM. Error can be restricted to a coarse level, which means shallower layers can provide better prior as a good guidance for deeper layers. In Supplementary Figure 4, we show the image difference between synthesized images and the ground truth. It can be observed that the proposed method yields synthesized images with remarkable visual similarity to reference images in lesion regions. Third, the stretch-out up-sampling module and the confidence map loss $\mathcal{L}_{\text{CM}}$ were also introduced into the proposed UCG-SAMR for the scenarios of unpaired data training. Thus, the practicality of the proposed methods is further demonstrated in the scenario where aligned pairs of lesion segmentation maps and multi-modal MR images are not required for training. As shown in Supplementary Figure 6, UCG-SAMR exhibits a similar performance for suppressing error in lesion regions. In addition, we separately evaluated the effectiveness of data augmentation by the synthesized data and the contribution to the synthesis quality for each of the designed modular network components in Section IV-E.

We adopted SSIM, and PSNR as evaluation criteria on holistic images, but these might not directly demonstrate diagnostic quality. To gain a better insight into the quality of synthesized lesion regions, we conducted evaluation of the pixel accuracy on three sub-regions: edema; cavity; and tumor, taking annotated lesion segmentation masks as the gold standard. In Table I, Shin et al. [33] slightly outperformed our method in SSIM and PSNR on holistic synthesized images, but CG-SAMR gained obvious improvement in lesion regions, increasing pixel accuracy by 13.6% for edema, 11.5% for cavity, and 17.4% for tumor. Leveraging the atlas of each sequence to provide brain anatomy prior rather than subject-specific WM, GM, and CSF masks may mainly contribute to richer diversity of the synthesized data and further benefit the data augmentation by synthesis (see Table II). Investigations of useful synthesis quality evaluation metrics that are close to a radiologist's judgment are still an ongoing topic in the medical image analysis community [77]. If such metrics can be used as a part of objective functions, the performance of synthesis networks could be improved further.

Under the scenario of using unpaired data, although our approach outperformed the baseline methods in all three synthesis metrics, SynSeg-Net [38] showed better segmentation performance in terms of dice score. It demonstrated the benefits of introducing extra segmentation networks based on the CycleGAN framework [52] and providing the direct supervision of manual segmentation masks. However, the segmentation performance of different methods is still not desirable under unpaired data training (Table IV). Notably, despite using the same metric with the BRATS challenge [21] (i.e. dice score), the segmentation labels and patient cohorts used in our study were different from those of BRATS challenge. Therefore, the difficulty of BRATS segmentation challenge [21] might not comparable with our segmentation task. In addition, the lack of direct supervision further compromised our segmentation performance in terms of training unpaired data. In future work to improve the segmentation performance of our UCG-SAMR model, the stretch-out sampling decoder for each label, over-complete representations segmentation architecture [78] and other adaptions will be added and evaluated.

While our proposed method yielded as convincing performance, there are several limitations in our current study. First, all subjects in this study were obtained from a single medical institution, so the deep-learning models were not trained, tested, or verified by outside data. This leads to a proportional bias in our study and a study-specific calibration is could not be performed. We have now begun to collect data from multiple external institutions to train and validate a generalizable algorithm. Second, our method is geared towards synthesizing 2D MR images. As shown in Section IV-A, the resolutions along the z-axis of APTw images is as large as four times that of the anatomic images (4.4mm v.s. 1.1mm). Thus, resampling APTw images to isotropic 3D volume results in a dramatic compromise of the fidelity of APTw images. Since isotropic data is indispensable for 3D convolution network, we adopted our approach in a 2D slice-based manner. Therefore, these two non-comparable resolutions limit the application of 3D methods. In our future work, we will acquire APTw images with high resolution, especially along the z-axis and hope the improved images will allow us to extend the proposed method to 3D synthesis. Third, the data augmentation by synthesis is only demonstrated on the lesion segmentation task. We believe that including various tasks to prove the quality of synthesized data would make this study more persuasive in the future work, but extra manual annotations will be required for training. Once the extra manual annotations are obtained, we can extend the proposed method to classification tasks. Fourth, we only conduct the confidence estimation at a scale of $\times 0.5$ (the half resolution). If we choose a smaller resolution (64×64), the feature maps would have contained more abstract features, which might cause undesirable intermediate synthesis results. If we choose a full resolution (256×256), the estimated confidence maps might not provide good guidance for the subsequent networks, since the full resolution feature maps approaches the end of the synthesis network. In addition, introducing more SM and CM modules on different scales might improve the final synthesis results. However, it also significantly increases the trainable parameters and subsequently increases the difficulty of training. Finally, our study leverages co-registered data, so the impact of misalignment between different MR sequences is ignored. Recent studies have shown that it is possible to increase robustness by introducing additional registration layers to correct the negative impacts from misalignment [29], [55]. Although a professional radiologist carefully monitored the image preprocessing in our study, it still would be nice to utilize these methods to mitigate the potential impairments from misalignment in our future work.

VI. Conclusion

In summary, we propose an effective generative model, called CG-SAMR, for multi-modal MR images, including

anatomic T_1w, Gd-T_1w, T_2w, and $FLAIR$, as well as molecular APTw MR images. It was shown that the proposed multi-task optimization under adversarial training further improved the synthesis quality in each ROI. The synthesized data could be used for data augmentation, particularly for images with pathological information about gliomas. Moreover, the proposed approach is an automatic, low-cost solution, which is capable of producing high quality data with diverse content that can be used to train data-driven methods. We further extend CG-SAMR to UCG-SAMR, demonstrating the feasibility of using unpaired data for training.

REFERENCES

[1] P. Y. Wen and S. Kesari, "Malignant gliomas in adults," *New England Journal of Medicine*, vol. 359, no. 5, pp. 492–507, 2008.

[2] J. K. Park *et al.*, "Scale to predict survival after surgery for recurrent glioblastoma multiforme," *Journal of clinical oncology*, vol. 28, no. 24, p. 3838, 2010.

[3] J. J. Vredenburgh *et al.*, "Bevacizumab plus irinotecan in recurrent glioblastoma multiforme," *Journal of clinical oncology*, vol. 25, no. 30, pp. 4722–4729, 2007.

[4] H. S. Birk, S. J. Han, and N. A. Butowski, "Treatment options for recurrent high-grade gliomas," *CNS oncology*, vol. 6, no. 1, pp. 61–70, 2017.

[5] P. Y. Wen *et al.*, "Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group," *Journal of clinical oncology*, vol. 28, no. 11, pp. 1963–1972, 2010.

[6] F. K. Albert, M. Forsting, K. Sartor, H.-P. Adams, and S. Kunze, "Early postoperative magnetic resonance imaging after resection of malignant glioma: objective evaluation of residual tumor and its influence on regrowth and prognosis," *Neurosurgery*, vol. 34, no. 1, pp. 45–61, 1994.

[7] J. Kalpathy-Cramer, E. R. Gerstner, K. E. Emblem, O. C. Andronesi, and B. Rosen, "Advanced magnetic resonance imaging of the physical processes in human glioblastoma," *Cancer research*, vol. 74, no. 17, pp. 4622–4637, 2014.

[8] J. Zhou, H.-Y. Heo, L. Knutsson, P. C. van Zijl, and S. Jiang, "Apt-weighted mri: Techniques, current neuro applications, and challenging issues," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 2, pp. 347–364, 2019.

[9] J. Zhou *et al.*, "Three-dimensional amide proton transfer mr imaging of gliomas: initial experience and comparison with gadolinium enhancement," *Journal of Magnetic Resonance Imaging*, vol. 38, no. 5, pp. 1119–1128, 2013.

[10] Y. S. Choi *et al.*, "Amide proton transfer imaging to discriminate between low-and high-grade gliomas: added value to apparent diffusion coefficient and relative cerebral blood volume," *European radiology*, vol. 27, no. 8, pp. 3181–3189, 2017.

[11] S. Jiang *et al.*, "Amide proton transfer-weighted magnetic resonance image-guided stereotactic biopsy in patients with newly diagnosed gliomas," *European Journal of Cancer*, vol. 83, pp. 9–18, 2017.

[12] S. Jiang, T. Zou, C. G. Eberhart, M. A. Villalobos, H.-Y. Heo *et al.*, "Predicting idh mutation status in grade ii gliomas using amide proton transfer-weighted (aptw) mri," *Magnetic resonance in medicine*, vol. 78, no. 3, pp. 1100–1109, 2017.

[13] B. Ma *et al.*, "Applying amide proton transfer-weighted mri to distinguish pseudoprogression from true progression in malignant gliomas," *Journal of Magnetic Resonance Imaging*, vol. 44, no. 2, pp. 456–462, 2016.

[14] N. Kumari, N. Thakur, H. R. Cho, and S. H. Choi, "Assessment of early therapeutic response to nitroxoline in temozolomide-resistant glioblastoma by amide proton transfer imaging: A preliminary comparative study with diffusion-weighted imaging," *Scientific reports*, vol. 9, no. 1, pp. 1–7, 2019.

[15] O. Togao *et al.*, "Amide proton transfer imaging of adult diffuse gliomas: correlation with histopathological grades," *Neuro-oncology*, vol. 16, no. 3, pp. 441–448, 2014.

[16] J. E. Park *et al.*, "Pre-and posttreatment glioma: comparison of amide proton transfer imaging with mr spectroscopy for biomarkers of tumor proliferation," *Radiology*, vol. 278, no. 2, pp. 514–523, 2016.

[17] S. Bisdas *et al.*, "Amide proton transfer mri can accurately stratify gliomas according to their idh mutation and 1p/19q co-deletion status." 2020.

[18] J. E. Park *et al.*, "Identification of early response to anti-angiogenic therapy in recurrent glioblastoma: Amide proton transfer–weighted and perfusion-weighted mri compared with diffusion-weighted mri," *Radiology*, vol. 295, no. 2, pp. 397–406, 2020.

[19] B. Joo *et al.*, "Amide proton transfer imaging might predict survival and idh mutation status in high-grade glioma," *European radiology*, vol. 29, no. 12, pp. 6643–6652, 2019.

[20] D. Paech *et al.*, "Relaxation-compensated amide proton transfer (apt) mri signal intensity is associated with survival and progression in high-grade glioma patients," *European radiology*, vol. 29, no. 9, pp. 4957–4967, 2019.

[21] 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*, 2018.

[22] P. Wang, P. Guo, J. Lu, J. Zhou, S. Jiang, and V. M. Patel, "Improving amide proton transfer-weighted mri reconstruction using t2-weighted images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 3–12.

[23] S. Tridandapani, "Radiology "hits refresh" with artificial intelligence," *Academic radiology*, vol. 25, no. 8, pp. 965–966, 2018.

[24] P. Guo, D. Li, and X. Li, "Deep oct image compression with convolutional neural networks," *Biomedical Optics Express*, vol. 11, no. 7, pp. 3543–3554, 2020.

[25] Y. Zhou, X. He, S. Cui, F. Zhu, L. Liu, and L. Shao, "High-resolution diabetic retinopathy image synthesis manipulated by grading and lesions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 505–513.

[26] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] T. Joyce, A. Chartsias, and S. A. Tsaftaris, "Robust multi-modal mr image synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 347–355.

[30] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 677–684.

[31] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal mr synthesis via modality-invariant latent representation," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 803–814, 2017.

[32] N. Cordier, H. Delingette, M. Lê, and N. Ayache, "Extended modality propagation: image synthesis of pathological cases," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2598–2608, 2016.

[33] H.-C. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*. Springer, 2018, pp. 1–11.

[34] R. Bala and R. Eschbach, "Spatial color-to-grayscale transform preserving chrominance edge information," vol. 2004, no. 1, pp. 82–86, 2004.

[35] N. Dai and F. Lee, "Edge effect analysis in a high-frequency transformer," in *Proceedings of 1994 Power Electronics Specialist Conference-PESC'94*, vol. 2. IEEE, 1994, pp. 850–855.

[36] P. Guo, P. Wang, J. Zhou, V. M. Patel, and S. Jiang, "Lesion mask-based simultaneous synthesis of anatomic and molecular mr images using a gan," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 104–113.

[37] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and synthesis in medical imaging," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 673–679, 2018.

[38] Y. Huo *et al.*, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–1025, 2019.

[39] S. Roy, A. Carass, and J. L. Prince, "Magnetic resonance image example-based contrast synthesis," *IEEE transactions on medical imaging*, vol. 32, no. 12, pp. 2348–2363, 2013.

[40] S. Roy, A. Carass, and J. Prince, "A compressed sensing approach for mr tissue contrast synthesis," in *Biennial International Conference on*

Information Processing in Medical Imaging. Springer, 2011, pp. 371–383.

[41] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical image analysis*, vol. 35, pp. 475–488, 2017.

[42] Y. Huang, L. Beltrachini, L. Shao, and A. F. Frangi, "Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2016, pp. 118–126.

[43] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander, "Mathematical textbook of deformable neuroanatomies," *Proceedings of the National Academy of Sciences*, vol. 90, no. 24, pp. 11 944–11 948, 1993.

[44] M. J. Cardoso, C. H. Sudre, M. Modat, and S. Ourselin, "Template-based multimodal joint generative model of brain data," in *International conference on information processing in medical imaging*. Springer, 2015, pp. 17–29.

[45] R. Li *et al.*, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 305–312.

[46] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 677–684.

[47] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2016, pp. 127–137.

[48] S. U. H. Dar, M. Özbey, A. B. Çatlı, and T. Çukur, "A transfer-learning approach for accelerated mri using deep neural networks," *Magnetic Resonance in Medicine*, vol. 84, no. 2, pp. 663–685, 2020.

[49] D. Nie *et al.*, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 417–425.

[50] H. R. Roth *et al.*, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical image analysis*, vol. 45, pp. 94–107, 2018.

[51] M. Yurt, S. U. H. Dar, A. Erdem, E. Erdem, and T. Çukur, "mustgan: Multi-stream generative adversarial networks for mr image synthesis," *arXiv preprint arXiv:1909.11504*, 2019.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[53] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[54] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Mr-to-ct synthesis using cycle-consistent generative adversarial networks," in *Proc. Neural Inf. Process. Syst.(NIPS)*, 2017.

[55] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[56] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *International workshop on simulation and synthesis in medical imaging*. Springer, 2017, pp. 3–13.

[57] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 9242–9251.

[58] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[59] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[60] R. Yasarla and V. M. Patel, "Confidence measure guided single image de-raining," *IEEE Transactions on Image Processing*, vol. 29, pp. 4544–4555, 2020.

[61] J. M. J. Valanarasu, R. Yasarla, P. Wang, I. Hacihaliloglu, and V. M. Patel, "Learning to segment brain anatomy from 2d ultrasound with less data," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[62] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[64] R. Yasarla and V. M. Patel, "Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8405–8414.

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

[66] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[67] S. Jiang *et al.*, "Identifying recurrent malignant glioma after treatment using amide proton transfer-weighted mr imaging: a validation study with image-guided stereotactic biopsy," *Clinical Cancer Research*, vol. 25, no. 2, pp. 552–561, 2019.

[68] Y. Zhang *et al.*, "Selecting the reference image for registration of cest series," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 3, pp. 756–761, 2016.

[69] J. Lipkova *et al.*, "Personalized radiotherapy design for glioblastoma: integrating mathematical tumor models, multimodal scans, and bayesian inference," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1875–1884, 2019.

[70] N. J. Tustison *et al.*, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[71] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[72] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[73] J. Ashburner *et al.*, "Spm12 manual," *Wellcome Trust Centre for Neuroimaging, London, UK*, p. 2464, 2014.

[74] T. Hinz, N. Navarro-Guerrero, S. Magg, and S. Wermter, "Speeding up the hyperparameter optimization of deep convolutional neural networks," *International Journal of Computational Intelligence and Applications*, vol. 17, no. 02, p. 1850008, 2018.

[75] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018.

[76] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 781–791, 2017.

[77] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[78] J. M. Jose, V. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 363–373.