Region-of-interest Attentive Heteromodal Variational Encoder-Decoder for Segmentation with Missing Modalities

Seung-wan Jeong[1,2], Hwan-ho Cho[3], Junmo Kwon[1,2], and Hyunjin Park[1,2(✉)]

[1] Sungkyunkwan University, Suwon, Republic of Korea
{jsw93,skenfn1231,hyunjinp}@skku.edu
[2] Center for Neuroscience Imaging Research, Suwon, Republic of Korea
[3] Konyang University, Daejon, Republic of Korea
hhcho@konyang.ac.kr

Abstract. The use of multimodal images generally improves segmentation. However, complete multimodal datasets are often unavailable due to clinical constraints. To address this problem, we propose a novel multimodal segmentation framework that is robust to missing modalities by using a region-of-interest (ROI) attentive modality completion. We use ROI attentive skip connection to focus on segmentation-related regions and a joint discriminator that combines tumor ROI attentive images and segmentation probability maps to learn segmentation-relevant shared latent representations. Our method is validated in the brain tumor segmentation challenge dataset of 285 cases for the three regions of the complete tumor, tumor core, and enhancing tumor. It is also validated on the ischemic stroke lesion segmentation challenge dataset with 28 cases of infarction lesions. Our method outperforms state-of-the-art methods in robust multimodal segmentation, achieving an average Dice of 84.15%, 75.59%, and 54.90% for the three types of brain tumor regions, respectively, and 48.29% for stroke lesions. Our method can improve the clinical workflow that requires multimodal images.

Keywords: Segmentation · Missing modalities · Multimodal learning · Adversarial learning

1 Introduction

Segmentation of lesions in medical images provides important information for assessing disease progression and surgical planning. Accurate segmentation often requires multimodal 3D images with complementary information about the lesions. For example, brain tumors are usually diagnosed with multimodal

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-26351-4_9.

L. Wang et al. (Eds.): ACCV 2022, LNCS 13846, pp. 132–148, 2023.
https://doi.org/10.1007/978-3-031-26351-4_9

magnetic resonance imaging (MRI) and different MRI modalities, such as T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid attenuation inversion recovery (FLAIR), provide complementary information (e.g., edema, enhancing tumor, and necrosis/non-enhancing tumor) about the brain tumor. In addition, T1, T2, diffusion-weighted images (DWI), and FLAIR MRI are acquired for the diagnosis of subacute ischemic stroke. DWI and FLAIR modalities provide general information about stroke lesions, whereas T1 and T2 modalities provide information on vasogenic edema present in subacute stroke [29]. Therefore, compared to the use of single-modality MRI, segmentation with multimodal MRI [11,13,14,16,18–21,27,30,40] helps to reduce uncertainty and improve segmentation performance.

Using multimodal data for segmentation is generally preferred, but modalities can often be missing in clinical practice. Some modalities may be missing due to limited patient tolerance, limited scan times, and corrupted images. In such cases, the missing modalities are not available for learning, which degrades the segmentation performance. Therefore, to fill the information gap of the missing modalities, an algorithm that effectively handles missing modalities is required. The simplest way to compensate for missing modalities is to impute missing modalities from other modalities using a method such as a k-nearest neighbor. However, this method cannot fully incorporate semantic information originally contained in the missing modalities. Many deep learning methods have recently been proposed to solve the problem of missing modalities [7,12,17,24,35,36,42]. These methods can be broadly grouped into two approaches. The first approach synthesizes missing modalities from available modalities and performs segmentation using complete modalities [24,36]. These methods are computationally complex because many different models are required to handle different missing scenarios. The second approach involves learning a shared representation of the multimodal information for segmentation. The learned shared representation is common to multimodal data, and thus, it is possible to construct one model that scales well to handle many missing scenarios.

Existing methods based on the second approach primarily use procedures to complete the full modalities as auxiliary tasks to learn a shared representation that is robust to missing modalities. Although this strategy successfully solves the problem of missing modalities, as a result, information about segmentation can be lost, which can lead to degradation of segmentation performance. Therefore, in addition to the constraints related to the completion of the full modalities, it is necessary to impose constraints related to segmentation tasks, such as image structure and the region-of-interest (ROI).

In this paper, we propose a new robust multimodal segmentation framework called region-of-interest attentive heteromodal variational encoder-decoder (RA-HVED). Our framework uses a heteromodal variational encoder-decoder (U-HVED) based on a multimodal variational formulation as a backbone to demonstrate the competitive performance of ROI attentive completion. The main contributions of our method are threefold: (1) We propose a robust segmentation framework for missing modalities that focuses on ROI. To impose

additional weights on the ROI, we introduce the ROI attentive skip connection module (RSM) and the ROI attentive joint discriminator. (2) We facilitate the learning of segmentation task-relevant shared representations by adding RSM that constrains the skip connection and an ROI attentive joint discriminator that strongly constrains modality completion. (3) We have conducted extensive experiments with missing modalities using brain tumor and stroke lesion datasets. Our method is more robust than previous methods for segmentation with missing modalities for the two datasets. In summary, our method can be applied to practical situations where data with missing modalities occur.

2 Related Works

2.1 Medical Image Synthesis

Medical image synthesis is a field that has recently been explored. Initially, methods based on convolutional neural network (CNN) have been commonly used for image synthesis. Li et al. [24] synthesized positron emission tomography (PET) images from MRI to improve the diagnosis of brain disease. Han [15] proposed a CNN method to synthesize the corresponding computed tomography (CT) images from an MRI. Since the first generation of CNNs, generative adversarial network (GAN)-based methods have achieved excellent performance in various medical image synthesis tasks. Nie et al. [31] synthesized CT images from MRI images using a context-aware GAN with high clinical utility. Costa et al. [10] generated a vessel tree using an adversarial autoencoder and synthesized a color retinal image from a vessel tree using a GAN. Wolterink et al. [38] used a GAN to obtain a routine-dose CT by reducing the noise of low-dose CT. Bi et al. [4] synthesized low-dose PET images from CT and tumor labels using a multichannel GAN. These methods are mostly intended for cases where one source modality is mapped to another target modality and thus are not suitable for multimodal settings where there may be more than one source and target modalities. Many studies on multimodal synthesis have recently been conducted to exploit the complementary information of multimodal data [5,23,34,37,41]. Wang et al. [37] synthesized full-dose PET images by combining low-dose PET images and multimodal MRI. Lee et al. [23] proposed CollaGAN for the imputation of missing image data. CollaGAN used a generator to produce a single output corresponding to each combination of multimodal inputs. CollaGAN used multiple cycle consistency to obtain the content of each combination, and the generation of the corresponding target modality was controlled by the one-hot-mask vector. However, this method cannot handle multiple missing modalities because it assumes that only one modality is missing at a time. Therefore, Shen et al. [34] proposed ReMIC for multiple missing modalities. Because ReMIC is a GAN framework that generates multiple images by separating the common content code of modalities and modality-specific style code, it can solve problems with missing multiple modalities. Furthermore, it has been shown that the learned content code contains semantic information and, therefore, can perform

segmentation tasks well. Because the segmentation task was performed independently after synthesis, the segmentation task was not explicitly optimized, and the segmentation performance depended on the results of the synthesized modalities. Therefore, we propose a robust segmentation framework for multiple missing modalities that overcomes these limitations.

2.2 Segmentation with Missing Modalities

Many methods have been proposed to solve the problem of the missing modality in segmentation [7,12,17,24,35,36,42]. These methods can be broadly divided into two types. The first approach synthesizes missing modalities and then performs segmentation from a set of complete modalities [24,36]. This approach can be effectively used when only two modalities are considered. However, once the number of modalities exceeds three, it becomes complex because many different models are required to handle different missing scenarios. Subsequently, the synthesis of the missing modalities in multimodal (more than two modalities) situations was proposed, but it is still difficult to deal with multiple missing modalities. The second approach involves creating a shared feature space that encodes multimodal segmentation information. Because this method finds common information via a shared encoder, it is possible to create one model that scales well to handle many missing scenarios. As such, many studies have adopted the second approach [7,12,17,35,42]. Havaei et al. [17] proposed a heteromodal image segmentation (HeMIS) method to calculate the statistics of learned feature maps for each modality and used them to predict the segmentation map. Because the encoder of HeMIS could not fully learn the shared representation using simple arithmetic operations, Dorent et al. [12] proposed U-HVED based on a multimodal variational formulation. U-HVED proved to be robust to missing modalities and outperformed the HeMIS in evaluating the brain tumor dataset. Chen et al. [7] applied the concept of feature disentanglement to effectively learn the shared latent representations in missing modality settings. However, this method requires an additional encoder for feature disentanglement. Shen et al. [35] introduced adversarial loss to learn invariant representations by matching feature maps of missing modalities to feature maps of complete modalities. This model was designed to be robust to only one missing modality; thus, it cannot handle situations where more than two modalities are missing. Existing methods [7,12,17,35,42] have proposed robust models for missing modalities using modality completion as an additional auxiliary task in the main segmentation task.

Our model goes further and improves the performance of segmentation with missing modalities by imposing constraints related to the segmentation task on the modality completion.

3 Methods

Figure 1 shows an overview of our proposed framework. As the backbone of our method, we first introduce U-HVED, which learns the multi-scale shared rep-

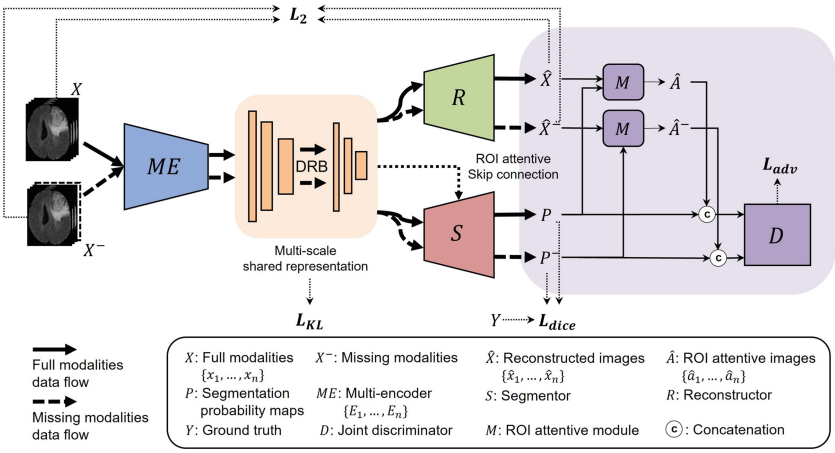

Fig. 1. Overview of our RA-HVED for robust multimodal segmentation. All modality-specific encoders $E_1, \ldots, E_n$ are included in the multi-encoder ME. The multi-scale shared representation from the multi-encoder ME flows into the reconstructor R for modality completion and segmentor S for image segmentation. The joint discriminator D computes adversarial loss using the concatenation $(\hat{A}, P)$ of the outputs from the two streams. ROI attentive images $\hat{A}$ are obtained using reconstructed images $\hat{X}$ and segmentation probability maps P.

resentation. This model extracts representations from encoders and fuses them into a shared representation of multimodal data. We also introduce a dimension reduction block (DRB) to efficiently learn the multi-scale shared representation. The shared representation is used in two streams for robust segmentation in different scenarios of missing modality. One stream generates the full modalities from the shared representation of the multimodal input, and the other performs the segmentation. At each level of the segmentor S, the encoder features are weighted by the segmentation-related regions using ROI attentive skip connections. Finally, we propose the ROI attentive module M and the joint discriminator D, which forces the reconstructor R to focus on the ROI.

3.1 Heteromodal Variational Encoder-Decoder

Dorent et al. [12] proposed U-HVED that combines U-net [33] and multimodal variational autoencoder (MVAE) [39] to perform segmentation from any subset of the full modalities. MVAE is a model developed in the context of conditionally independent modalities $X = x_1, \ldots, x_n$ when a common latent variable z is given. The authors of MVAE deal with the missing data by extending variational autoencoder (VAE) [22] formulation for multimodal inputs. The encoded mean μ and covariance Σ of each modality are fused into a common latent variable z of the multimodal data using the product of Gaussian (PoG) [6] (Supplementary Fig. 2(a)). If a modality is missing, the corresponding variation parameters are

excluded (Supplementary Fig. 2(b)). The latent variable z estimated by sampling was decoded into the image space. Sampling was performed using a reparameterization trick [22]. U-HVED performs optimization by drawing random subsets in each iteration. VAE loss for the network optimization is as follows:

$$L_{VAE} = \mathbb{E}_{x^-}[D_{KL}[ME(x^-) \parallel N(0,1)] + \parallel \hat{x}^- - x \parallel] \ , \tag{1}$$

where x^- are the random missing modalities from input images x, $\hat{x}^-$ denotes the reconstructed images, $D_K L$ is KL divergence, ME is a multi-encoder $E_1, \ldots, E_n$, and $N(0, I)$ is the normal distribution.

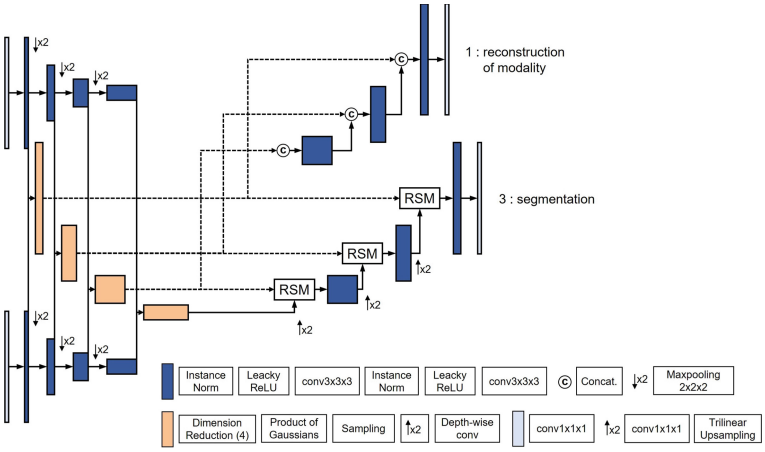

Fig. 2. Schematic visualization of our network architecture. Only two encoders, a segmentation decoder and a reconstruction decoder are shown. Each orange block stands for DRB. The size of the output channel of the decoder is 1 for the reconstruction of modality and 3 for segmentation. (Color figure online)

3.2 ROI Attentive Skip Connection Module

U-net [33] uses skip connections for successful segmentation. Generally, skip connection of U-net is a structure in which an input of a decoder and a feature of a corresponding encoder are concatenated. Since the decoder uses the encoder's features, it is easier to recover the lost detailed spatial information. Here, we propose an ROI attentive skip connection module (RSM) to emphasize the segmentation-related region in encoder features. Before applying RSM, a dimension reduction block for efficient representation learning is introduced.

Dimension Reduction Block. U-HVED learns multi-scale shared representation by applying MVAE to skip connections and the main stream of U-net. As

the layer depth increases, the dimension of the representation increases, which makes learning the shared representation difficult. This problem is magnified in 3D medical images compared to 2D natural images because a two-fold magnification of the image leads to an eight-fold increase in the amount of data in 3D compared to the four-fold increase in 2D images. As the spatial size of the shared representation eventually increases, the expansion of the model becomes limited. To solve this problem, we propose DRB method that reduces the dimensions of the shared representation. DRB consists of dimension reduction and upsampling (Fig. 2). First, our DRB reduces the size of the spatial and channel by a $3 \times 3 \times 3$ convolutional layer. Then, after sampling the representation, the representation is restored to the original dimension by a $1 \times 1 \times 1$ convolutional layer, an upsampling layer, and depth-wise convolutional layer. DRB is applied to each modality at all levels. Ultimately, we obtain a shared representation with an 8-fold reduction in spatial size and a 2-fold reduction in channel size compared to the original U-HVED. This shared representation has a relatively small dimension compared to the original dimension, which enables efficient learning and facilitates 3D expansion of the model.

ROI Attentive Skip Connection. Our RSM does not simply concatenate input features when applying skip connection, but applies weights to encoder features using segmentation feature maps and then proceeds with concatenation. In Fig. 3, spatial features are obtained by using channel-wise average and max operations of the l-th segmentation feature f_l^S and l-th encoder features f_l^E. The concatenated spatial features are transformed into spatial attention weights by sequentially passing them through the depth-wise convolutional layer, the point-wise convolutional layer, and the sigmoid activation. Spatial attention weights are applied to the encoder features in a residual manner. The spatial attention weights of segmentation features f_l^S are obtained through the same process with only their own spatial features and are applied to the segmentation features f_l^S in a residual manner. Finally, the attentive segmentation and encoder features are concatenated.

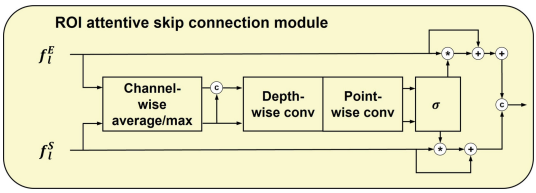

Fig. 3. Schematic visualization of our ROI attentive skip connection module. Using f_l^S, spatial attention weights are applied to each feature.

3.3 ROI Attentive Module and Joint Discriminator

Many studies [1,7,8,12,17,28,30,32,35,42] used an autoencoder or VAE as an auxiliary task to learn meaningful representations for segmentation. These methods have achieved success in accurate segmentation. However, the learned shared representation may be less relevant for segmentation because the network is simultaneously trained for image reconstruction and segmentation. Therefore, we introduce a joint discriminator D that combines ROI attentive images $\hat{a}$ and segmentation probability maps p to learn a shared representation that focuses on the segmentation task. The joint discriminator D enables learning of the shared representation by imposing additional constraints on the image reconstruction. ROI attentive images $\hat{a}$ are created by the ROI attentive module M (Eq. 2) using the reconstructed images $\hat{x}$ and their segmentation probability maps p as inputs. These values are calculated as follows:

$$\hat{a} = \hat{x} * (1 + \sum_{k} p_k) \ , \tag{2}$$

where p_k is the segmentation probability maps whose values are greater than 0.5 for segmentation class k. The ROI attentive module M emphasizes the ROI by using the average of the segmentation probability maps p in the reconstructed images $\hat{x}$ as a weight. Joint discriminator D is trained as an adversary by distinguishing between full and missing modalities with a focus on the ROI. The adversarial loss for joint discriminator D is defined as

$$L_{adv} = \mathbb{E}_{\hat{a}}[\log(D(\hat{a}, p))] + \mathbb{E}_{\hat{a}^-}[\log(D(\hat{a}^-, p^-))] \ . \tag{3}$$

Although it is possible to constrain the ROI in image reconstruction using only the joint discriminator D, we enforce stronger constraints on the ROI through the ROI attentive module M. Thus, our joint discriminator D strongly constrains the reconstruction network R to reconstruct images that focus on the ROI, making the shared representation more relevant to the segmentation task and more robust to missing modalities.

3.4 Segmentation

We choose a Dice loss for segmentation network $ME \circ S$ that consists of multi-encoder ME and segmentor S. Our goal is to successfully perform segmentation in all subsets of input modalities; thus, we use both Dice loss for full modalities x and missing modalities x^- to train a segmentation network $ME \circ S$.

$$L_{seg} = Dice(y, p) + Dice(y, p^-) \ , \tag{4}$$

where y is ground truth, and p represents is segmentation probability maps.

3.5 Training Process

As shown in Fig. 1, our goal is to learn a multi-scale shared representation for multiple encoders. In this context, segmentor S and reconstructor R are trained for segmentation and modality completion, respectively, using a multi-scale shared representation. Finally, through joint adversarial learning, segmentor S is forced to generate a segmentation map that is robust to missing modalities, and the reconstructor R is forced to generate images related to the segmentation task. The total objective function with trade-off parameters λ_1,λ_1 for the entire framework is as follows:

$$L = L_{seg} + \lambda_1 * L_{VAE} + \lambda_2 * L_{adv} \ . \tag{5}$$

4 Experiments

4.1 Data

BraTS. We evaluated our method using a multimodal brain tumor segmentation challenge (BraTS) 2018 dataset [2,3,26]. The imaging dataset included T1, T1ce, T2, and FLAIR MRI modalities for 285 patients with 210 high-grade gliomas and 75 low-grade gliomas. Imaging data were preprocessed by resampling to an isotropic 1 mm3 resolution, coregistration onto an anatomical template, and skull stripping. The ground truth of the three tumor regions was provided by manual labeling by experts. The clinical goal of BraTS 2018 dataset is to segment three overlapping regions (*i.e.*, complete tumor, tumor core, and enhancing tumor). We randomly divided the dataset into 70 % training, 10 % validation, and 20 % testing sets.

ISLES. The ischemic stroke lesion segmentation challenge (ISLES) 2015 dataset [25] provides multimodal MRI data. We selected the subacute ischemic stroke lesion segmentation (SISS) task between the two subtasks. The SISS dataset provides four MRI modalities consisting of T1, T2, DWI, and FLAIR for 28 patients. The imaging data were preprocessed by resampling to an isotropic 1 mm3 resolution, coregistration onto the FLAIR, and skull stripping. The infarcted areas were manually labeled by the experts. We randomly divided the dataset into 70 % training and 30 % testing sets.

4.2 Implementation Details

The network structure of RA-HVED is shown in Fig. 2. Our entire network takes the form of a 3D U-net [9]. A detailed network structure is referred in the supplementary material. We normalized the MRI intensity to zero mean and unit variance for the whole brain in each MRI modality. For data augmentation, we randomly applied an intensity shift for each modality and flipped for all axes. The 3D images were randomly cropped into 112 × 112 × 112 patches

and used during training. We used an Adam optimizer with an initial learning rate of 1e-4 and a batch size of 1. The learning rate was multiplied by (1-epoch/total_epoch)$^{0.9}$ for each epoch during 360 epochs. We set $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$ through a grid search with 0.1 increments in [0, 1] from the validation set. The missing modalities were constructed by uniformly drawing subsets, as is done in U-HVED [12] during training. All networks were implemented using the Pytorch library. To obtain the final inference result, we used the sliding window strategy, which is commonly used in patch-based networks. The stride of the sliding window is $28 \times 28 \times 28$ and equal to patch size / 4. After we obtained the final result, no postprocessing was done. The settings of the brain tumor dataset were used in the stroke dataset without any specific architectural changes or hyperparameter tuning. Our implementation is available here.[1]

4.3 Results of Segmentation with Missing Modalities

To evaluate the robustness of our method against the missing modalities, we compare our method (RA-HVED) to three previous methods in all scenarios of missing modalities: 1) U-HeMIS [12] is a U-net variant of HeMIS [17]-a first model for learning shared representations for missing modalities. 2) U-HVED [12], which combines MVAE and U-net, is compared because it is the base of

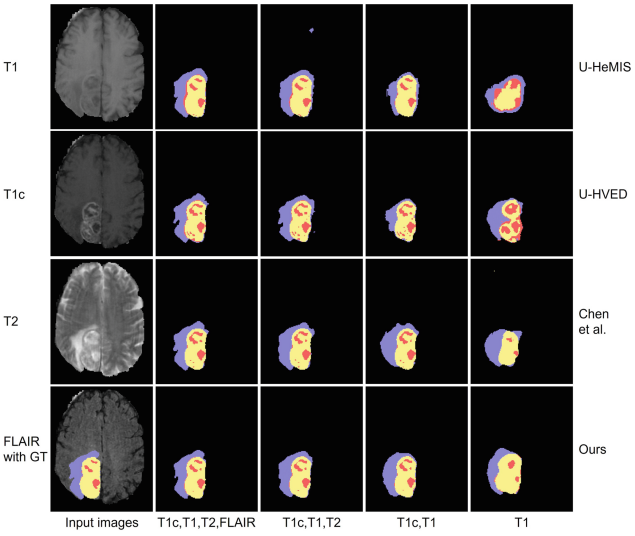

Fig. 4. Results of tumor segmentation produced by using different methods on the BraTS dataset. The first column is the input image of each modality, and each row shows segmentation results with missing modalities of comparison methods. GT: ground truth. Purple: complete tumor; Yellow: tumor core; Red: enhancing tumor. (Color figure online)

[1] https://github.com/ssjx10.

our method. 3) Chen et al. [7] is compared due to the strength of feature disentanglement in learning shared representations. It separates each modality into content and appearance codes. Then, segmentation is performed using the shared representation created by fusing the content codes.

Table 1. Comparison of segmentation performance with respect to all 15 missing modality scenarios on the BraTS 2018 dataset. The presence of modality is denoted by •, and the missing of modality is denoted by ◦. All results are evaluated with a dice score (%).

Available Modalities				Complete tumor				Tumor core				Enhancing tumor			
F	T1	T1c	T2	U- HeMIS	U- HVED	Chen et al.	Ours	U- HeMIS	U- HVED	Chen et al.	Ours	U- HeMIS	U- HVED	Chen et al.	Ours
◦	◦	◦	•	79.00	78.89	80.60	**81.47**	60.95	67.34	60.60	**67.82**	27.54	34.97	30.15	**36.93**
◦	◦	•	◦	57.86	58.26	68.92	**71.83**	66.95	71.99	75.57	**77.93**	59.68	61.35	65.97	**68.97**
◦	•	◦	◦	59.40	**68.38**	67.79	68.16	46.49	**58.60**	50.28	57.46	16.68	22.35	23.54	**28.75**
•	◦	◦	◦	84.57	82.72	85.04	**88.21**	63.54	60.33	63.18	**65.62**	27.42	29.89	29.06	**36.96**
◦	◦	•	•	81.16	79.85	83.76	**84.63**	76.78	80.66	81.64	**81.68**	64.74	66.94	68.16	**70.29**
◦	•	•	◦	67.44	73.06	72.96	**75.27**	68.03	78.06	78.41	**80.35**	64.08	66.84	66.62	**69.97**
•	•	◦	◦	86.08	84.94	87.17	**88.84**	66.00	67.66	67.86	**70.73**	32.24	31.84	34.66	**38.12**
◦	•	◦	•	80.47	82.32	83.77	**83.97**	62.73	**72.76**	64.33	71.59	31.26	36.93	34.75	**39.19**
•	◦	◦	•	86.77	88.38	87.63	**88.94**	67.04	**71.94**	65.76	70.98	33.89	**40.69**	33.67	40.32
•	◦	•	◦	86.67	85.48	87.08	**89.45**	76.64	76.13	82.14	**83.17**	64.71	63.14	67.96	**69.76**
•	•	•	◦	86.54	86.78	87.96	**89.12**	77.58	77.69	83.20	**84.14**	65.65	65.45	68.67	**70.87**
•	•	◦	•	87.19	**88.59**	88.31	88.47	67.03	73.15	67.71	**73.58**	35.86	40.31	37.65	**41.71**
•	◦	•	•	87.32	88.31	88.15	**89.31**	78.31	80.44	82.42	**82.65**	66.45	67.75	68.54	**69.89**
◦	•	•	•	81.58	82.04	84.16	**84.89**	75.84	81.39	82.29	**82.59**	66.16	68.40	68.07	**70.54**
•	•	•	•	87.56	88.10	88.50	**89.64**	78.32	82.30	83.27	**83.62**	66.50	68.31	68.31	**71.28**
Average				79.97	81.07	82.79	**84.15**	68.82	73.36	72.58	**75.59**	48.19	51.01	51.05	**54.90**

Results of BraTS. Table 1 shows the brain tumor segmentation results for various methods to deal with missing modalities. Our method outperforms the segmentation accuracy of previous methods for all three tumor regions in most missing modality scenarios. Our method achieved the highest average Dice of 84.15%, 75.59%, and 54.90% for the three nested tumor regions. The second robust method is Chen's approach, which achieves an average Dice of 82.79%, 72.58%, and 51.05%. We show that the Dice score increases more in the case of enhancing tumors than in other tumor regions. FLAIR and T2 modalities provide valuable information for complete tumors, and T1c modality provides crucial information for tumor cores and enhancing tumors. Because the T1 modality has relatively little information about the tumor compared to other modalities, it is difficult to obtain robust results when only the T1 modality is available. However, our method achieves similar or even higher accuracy than U-HVED, which shows high performance even in the case of the T1 modality alone. This indicates that the proposed method successfully learns shared representations. Inference times for other methods are referred in the supplementary Table 1.

Figure 4 shows a qualitative comparison of the various methods. As the number of missing modalities increases, the segmentation results of all methods gradually deteriorate. Nevertheless, our method provides more robust segmentation results than other methods and achieves accurate segmentation results even for the T1 modality, which contains relatively little information.

Table 2. Comparison of segmentation performance on the ISLES 2015 dataset.

Methods	U-HeMIS	U-HVED	Chen et al.	Ours
Average Dice score (%)	40.09	42.92	41.16	**48.29**

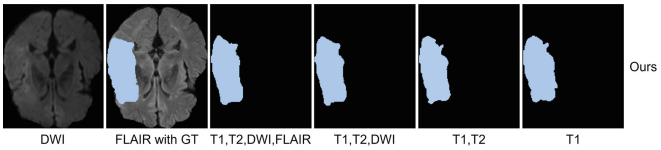

Fig. 5. Results of stroke lesion segmentation produced by our RA-HVED on ISLES.

Results of ISLES. The segmentation results for ISLES are shown in Table 2. On the average Dice score, our method shows a higher segmentation accuracy than other methods, reaching 48.29%. Chen's approach achieves a lower segmentation accuracy than U-HVED, in contrast to the results of the BraTS dataset. Figure 5 shows the results of stroke lesion segmentation using our method. Even when the number of missing modalities increases, our method provides robust segmentation results. Segmentation results for all missing scenarios and visualization of segmentation about other methods are referred in the supplementary material.

4.4 Results of Reconstruction with Missing Modalities

Our primary goal is to perform segmentation, but reconstruction of modalities can be performed during the process. Figure 6 shows the results of image reconstruction on FLAIR, the modality with the most information, for BraTS when modalities are missing. When all modalities are available, U-HeMIS and U-HVED produce images that are most like the corresponding image. However, other methods, including U-HVED, fail to produce tumor area details when the number of missing modalities increases. When only the T1 modality is available, the details of the tumor core are poorly generated. When all modalities are available, our method generates images similar to manual segmentation, although it is less similar to the corresponding image for the tumor region. Moreover, our method generates details of the tumor core better than other methods, even when the number of missing modalities increases. The reconstruction result for ISLES is referred in the supplementary material.

4.5 Ablation Study

We conduct an ablation study on RA-HVED with U-HVED as the baseline. In Table 3, we compare the methods using the average Dice score for all possible subsets of input modalities on the BraTS dataset. First, we confirm the effect

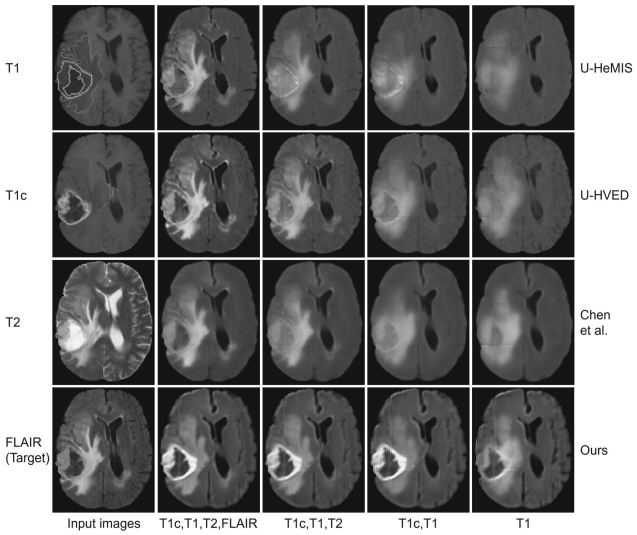

Fig. 6. Image reconstruction results generated by different methods on the BraTS dataset. The first column is the input image of each modality, and each row shows the reconstruction results with missing modalities of the comparison methods. Ground truth for segmentation is overlaid on T1. Purple: complete tumor; Yellow: tumor core; Red: enhancing tumor. (Color figure online)

Table 3. Ablation study of our key components. DRB : dimension reduction block, RSM : ROI attentive skip connection module, RJD : ROI attentive joint discriminator.

Methods	**Average Dice Score (%)**		
	Complete tumor	Tumor core	Enhancing tumor
(1) U-HVED	81.07	73.36	51.01
(2) U-HVED + DRB	81.34	73.03	51.36
(3) (2) + RSM	82.68	74.27	53.73
(4) (2) + RSM + RJD (RA-HVED)	**84.15**	**75.59**	**54.90**

of adding DRB to U-HVED when comparing (1) with (2). In Method (2), the dimension of the shared representation is decreased compared to (1), but the average Dice scores are similar. In method (3), RSM improves overall segmentation performance including enhancing tumor. In particular, the segmentation performance in enhancing tumor region is further improved because attention is imposed using segmentation features. Finally, in (4), an ROI attentive joint discriminator is added to (3) to provide stronger constraints to the ROI in the image reconstruction. The ROI attentive module is added to improve the segmentation performance in all tumor regions and achieve the highest Dice score in most scenarios with missing modalities. In particular, the average Dice increases by 4.5%

Table 4. Results on the effectiveness of ROI-based attention in RSM and ROI attentive joint discriminator.

Methods	**Average Dice Score (%)**		
	Complete tumor	Tumor core	Enhancing tumor
spatial-wise attention	83.16	74.24	53.69
joint discriminator	83.28	74.61	54.14
Ours (RA-HVED)	**84.15**	**75.59**	**54.90**

for (2) in the enhancing tumor. This shows that the proposed key components of RA-HVED efficiently learn the multi-scale shared representations.

In Table 4, we conduct experiments to prove the effect of ROI-based attention in RSM and ROI attentive joint discriminator. First, spatial-wise attention is applied to encoder features without the intervention of segmentation features in RSM (spatial attention in Table 4). Next, the joint discriminator replaces the ROI attentive joint discriminator (joint discriminator in Table 4). Both models achieve lower segmentation performance than RA-HVED. This indicates that ROI-based attention is important for learning segmentation-relevant shared representations.

5 Conclusion

In this study, we propose a novel and robust multimodal segmentation method that can function effectively when there are missing modalities. Our model efficiently learns segmentation-relevant shared representations through ROI attentive skip connection and joint adversarial learning that constrains the ROI in modality completion. We validate our method on a brain tumor and a stroke lesion dataset. Experimental results show that the proposed method outperforms previous segmentation methods on missing modalities. Moreover, we demonstrate the effectiveness of our key components in an ablation study. Our method can be applied to improve the clinical workflow that requires multimodal images.

Acknowledgements. This research was supported by the National Research Foundation (NRF-2020M3E5D2A01084892), Institute for Basic Science (IBS-R015-D1), Ministry of Science and ICT (IITP-2020-2018-0-01798), AI Graduate School Support Program (2019-0-00421), ICT Creative Consilience program (IITP-2020-0-01821), and Artificial Intelligence Innovation Hub (2021-0-02068).

References

1. Amyar, A., Modzelewski, R., Li, H., Ruan, S.: Multi-task deep learning based CT imaging analysis for Covid-19 pneumonia: classification and segmentation. Comput. Biol. Med. **126**, 104037 (2020)

2. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(1), 1–13 (2017)
3. Bakas, S., 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)
4. Bi, L., Kim, J., Kumar, A., Feng, D., Fulham, M.: Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs). In: Cardoso, M.J., et al. (eds.) CMMI/SWITCH/RAMBO -2017. LNCS, vol. 10555, pp. 43–51. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67564-0_5
5. Cao, B., Zhang, H., Wang, N., Gao, X., Shen, D.: Auto-GAN: self-supervised collaborative learning for medical image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10486–10493 (2020)
6. Cao, Y., Fleet, D.J.: Generalized product of experts for automatic and principled fusion of gaussian process predictions. arXiv preprint arXiv:1410.7827 (2014)
7. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_50
8. Chen, S., Bortsova, G., García-Uceda Juárez, A., van Tulder, G., de Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 457–465. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_51
9. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
10. Costa, P., et al.: End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imaging **37**(3), 781–791 (2017)
11. Cui, S., Mao, L., Jiang, J., Liu, C., Xiong, S.: Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network. Journal of healthcare engineering 2018 (2018)
12. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_9
13. Feng, C., Zhao, D., Huang, M.: Segmentation of ischemic stroke lesions in multi-spectral MR images using weighting suppressed FCM and three phase level set. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 233–245. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_20
14. Halme, H.-L., Korvenoja, A., Salli, E.: ISLES (SISS) Challenge 2015: segmentation of stroke lesions using spatial normalization, random forest classification and contextual clustering. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 211–221. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_18
15. Han, X.: MR-based synthetic CT generation using a deep convolutional neural network method. Med. Phys. **44**(4), 1408–1419 (2017)
16. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)

17. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: hetero-modal image segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_54
18. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnU-net for brain tumor segmentation. In: Crimi, A., Bakas, S. (eds.) BrainLes 2020. LNCS, vol. 12659, pp. 118–132. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72087-2_11
19. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
20. Jiang, Z., Ding, C., Liu, M., Tao, D.: Two-Stage Cascaded U-Net: 1st place solution to BraTS challenge 2019 segmentation task. In: Crimi, A., Bakas, S. (eds.) BrainLes 2019. LNCS, vol. 11992, pp. 231–241. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46640-4_22
21. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Lee, D., Kim, J., Moon, W.J., Ye, J.C.: CollaGAN: collaborative GAN for missing image data imputation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2019)
24. Li, R., et al.: Deep learning based imaging data completion for improved brain disease diagnosis. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8675, pp. 305–312. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10443-0_39
25. Maier, O., et al.: Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. **35**, 250–269 (2017)
26. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
27. Mitra, J., et al.: Lesion segmentation from multimodal MRI using random forest following ischemic stroke. Neuroimage **98**, 324–335 (2014)
28. Moeskops, P., et al.: Deep learning for multi-task medical image segmentation in multiple modalities. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 478–486. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_55
29. Muir, K.W., Buchan, A., von Kummer, R., Rother, J., Baron, J.C.: Imaging of acute stroke. Lancet Neurol. **5**(9), 755–768 (2006)
30. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
31. Nie, D., et al.: Medical image synthesis with context-aware generative adversarial networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 417–425. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_48
32. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 669–677. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_74

33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
34. Shen, L., et al.: Multi-domain image completion for random missing input data. IEEE Trans. Med. Imaging **40**(4), 1113–1122 (2020)
35. Shen, Y., Gao, M.: Brain tumor segmentation on MRI with missing modalities. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 417–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_32
36. van Tulder, G., de Bruijne, M.: Why does synthesized data improve multi-sequence classification? In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 531–538. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_65
37. Wang, Y., et al.: 3D auto-context-based locality adaptive multi-modality GANs for pet synthesis. IEEE Trans. Med. Imaging **38**(6), 1328–1339 (2018)
38. Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans. Med. Imaging **36**(12), 2536–2545 (2017)
39. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
40. Zhou, C., Ding, C., Wang, X., Lu, Z., Tao, D.: One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. IEEE Trans. Image Process. **29**, 4516–4529 (2020)
41. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-Net: hybrid-fusion network for multi-modal MR image synthesis. IEEE Trans. Med. Imaging **39**(9), 2772–2781 (2020)
42. Zhou, T., Canu, S., Vera, P., Ruan, S.: Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 533–541. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_52