A Unified Missing Modality Imputation Model with Inter-Modality Contrastive and Consistent Learning

Liangce Qi[1], Yusi Liu[1], Yuqin Li[1], Weili Shi[1], Guanyuan Feng[1], Zhengang Jiang[1,*]

[1]Changchun University of Science and Technology, Changchun, Jilin, China
jiangzhengang@cust.edu.cn

Abstract. Multi-modality magnetic resonance imaging (MRI) is widely used in the clinical diagnosis of brain tumors. However, the issue of missing modalities is frequently encountered in the real-world setting and can lead to the collapse of deep-learning-based automatic diagnosis algorithms that rely on full-modality images. To address this challenge, we propose a unified model capable of synthesizing missing modalities through any subsets of the full-modality images. Our method is a sequence-to-sequence prediction model that predicts the missing images by inter-modality correlation and modality-specific semantics. Specifically, we develop a dual-branch encoder, where both branches encode partially masked image tokens into low-dimensional features independently. A decoder then generates the target input images based on the fused encoder features. To strengthen the representative ability of encoder features, we propose a combination loss to improve the discriminative and consistency between diverse modality features. We evaluate our method on the BraTS 2023 dataset. Extensive quantitative and qualitative experiments demonstrate the high fidelity and utility of the synthesized images.

Keywords: Missing image imputation · Multi-modality MRI · Brain tumor segmentation.

1 Introduction

Gliomas are a prevalent type of brain tumor that poses a significant threat to patient well-being and quality of life. In clinical practice, multi-modality MRI is the most commonly used diagnostic tool because it provides complementary information about the structure and tissue of the brain. Typical used modalities include T1-weighted (T_1), post gadolinium contrast T1-weighted (T_{1c}), T2-weighted (T_2), and T2 Fluid Attenuated Inversion Recovery (T_f). Nowadays, deep-learning based segmentation methods [12, 3] have achieved promising performance in detecting brain tumors based on full-modality MRI images. However, due to problems such as scanning time and cost, clinical diagnosis often faces the issue of missing modalities, which greatly reduces the accuracy of these algorithms. For example, D^2-Net [16] points out that using four modality images can achieve a Dice score of 82.19% on the BraTS 2018 dataset [9], while the best results using a single modality degrade to 44.97%. Therefore, the new approach that can overcome this challenge is attractive in clinical scenarios.

There are currently two mainstream solutions. The first is developing segmentation models that can adapt to various missing modalities situation [16, 10, 18]. One common idea of these approaches is to learn multiple modality-incomplete features during training and combine them to reconstruct high representational modality-complete features for segmentation during inference. Some other research [19, 7, 1, 6] involves imputing the missing modality images. The generation results can be used for both manual and automatic diagnosis. In comparison, these methods have better interpretability and wider applications. Early attempts [20, 17] formulated this task as a style transfer or generation task, with the input and output modalities being fixed. One obvious problem of these methods is the necessity to train multiple models to encompass all input-output combinations of total m modalities, eg., $2^m - 2$ when using a one-to-one model. Therefore, the unified model capable of accommodating any modality combination and generating missing modalities has garnered more attention.

To build a unified model, it is first necessary to address the challenge of unfixed input and output length. One intuitive solution is to build independent encoders to handle different modalities [19, 13]. Another is developing a single encoder-decoder model based on the self-attention mechanism [15] and formulating the missing modality imputation task into the sequence prediction task. The former methods can achieve better generation quality, especially when fewer modalities are available. The latter is more straightforward in preserving the anatomical information among multi-modality images. Both methods also developed various modules to enhance the generation quality, typically including the entanglement and disentanglement of modality-invariant features and modality-specific features from the modality-complete features. In general, these methods obtain the input features through inter-modality relations, while the generation of each modality is still independent. In contrast, we introduce the inter-modality constraint into the generation process to improve the overall generation quality.

In this paper, we propose a self-attention-based unified missing modality imputation model that is capable of reconstructing four MRI modalities from any subset of itself. We show that ensuring multi-modality images are equally discriminative and relevant in feature space as in pixel space is important to improve the quality of the generated images. To this end, we develop a dual-branch encoder and single-decoder network architecture that leverages the correlation between modalities and semantic information within the image to predict the missing image. The imputation model is trained by proposed inter-modality contrastive and consistent learning. The effectiveness of the proposed method is evaluated on the BraTS 2023 dataset [9]. Extensive quantitative and qualitative experiments demonstrated the high quality and utility of synthesized images. Our codes are available at https://github.com/LcQi-mic/mod_imp.

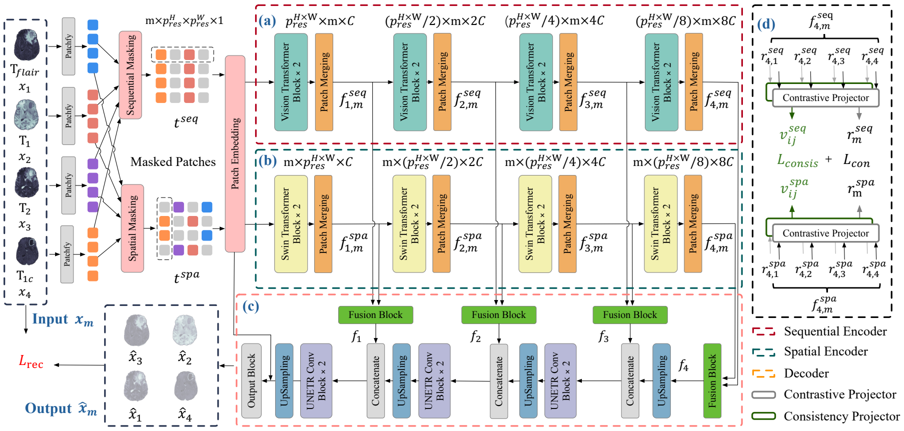

Fig. 1. Overview of the proposed approach. (a) The sequential encoder aims to obtain inter-modality correlations. (b) The spatial encoder aims to obtain semantics within images. (c) Decoder used to reconstruct missing images. (d) Inter-modality contrastive and consistent learning.

2 Method

The proposed method employs a dual-branch encoder and one decoder configuration, with the overview architecture illustrated in Fig. 1. The input and output of generation backbone are four modality MRI images denoted as $x = \{x_m, m \in [1, 2, 3, 4]\} \in R^{H\times W}$ and $\hat{x} = \{\hat{x}_m\} \in R^{H\times W}$ with $H \times W$ being the size of the 2D images. After translating x_m into partially masked image patches, the sequential masking layer simulates the modality missing by randomly masking all patches belonging to 1 to $m - 1$ modalities to ensure at least 1 modality are maintained. Meanwhile, the spatial masking layer randomly masks patches of each modality with a high masking ratio, e.g., 75% in this paper. Then, one shared patch embedding layer maps two masked patches into image tokens $(t^{seq}, t^{spa}) \in R^{m\times p_{res}^{H} \times p_{res}^{W} \times C}$, where $p_{res}^{H\times W} = \frac{H\times W}{p_{size}}$ is the patch resolution and $p_{size} = 2 \times 2$ is the patch size. The sequential and spatial branch receive t^{seq} and t^{spa} to obtain low-dimensional features $f_{i,m}^{seq}$ and $f_{i,m}^{spa}$, i is the index of encoding stage. Finally, a decoder reconstructs the original images $\hat{x}_m$ based on the fusion features f_i. When inference, the input of two branches are the same t^{seq}.

2.1 Architecture design

The encoder uses a parallel network architecture as seen in Fig. 1 (a, b). The input tokens are reshaped and forwarded through each branch simultaneously. We have a stack of encoding blocks in both encoder branches as well as the decoder. Each block in the encoder consists of two transformer blocks to model constant-range dependencies and follow one patch merging layer to reduce the spatial resolution and increase the number of channels by a factor of two. Similar to U-Net [12], the encoder and decoder are skip-connected at the same encoding stage.

The sequential branch (Fig. 1 a) predicts the missing patch features by learning the correlation information among tokens belonging to different modalities at the same spatial location. The input sequence length is m and the transformer block is adopted from the vanilla vision transformer [2] with absolute position embedding. The spatial branch (Fig. 1 b) learns the semantics within each modality image following the masked image modeling [5] paradigm. The input sequence length is $p_{res}^{H\times W}$ and the transformer block is adopted from Swin Transformer [8] with relative position embedding. Another goal of the spatial branch is to provide constraints on obtained modality features which will be introduced in the next subsection.

To combine two incomplete and complementary branch features, we add a fusion block after the patch merging layer in each stage. Specifically, the fusion block first adds two input features, then sequentially passes them into a spatial attention module M_{spa} and a channel attention module M_{cha} to get the final features f_i. For convenience, given an input feature $f_i^{'} = f_i^{spa} + f_i^{seq}$, the overall process can be summarised as $f_i = M_{cha}(M_{spa}(f_i^{'}))$. The spatial attention is computed as:

$$M_{spa}(f_i^{'}) = Sigmoid(Conv^{1*1}(Conv^{3*3}(f_i^{'}) + Conv^{7*7}(f_i^{'}))) \times f_i^{'} \tag{1}$$

where $Conv^{n\times n}$ indicates a convolutional layer with n kernel size. Given input feature $f_i^{''}$, the channel attention is computed as:

$$M_{cha}(f_i^{''}) = Sigmoid(ReLU(MLP(MaxPool(f_i^{''})))) \times f_i^{''} \tag{2}$$

The convolutional block in the decoder (Fig. 1 c) is adopted from UNETR [4]. The output features are fed into four output blocks to generate images, each is a 1×1 convolutional layer following a LeakyReLU activation function. More details can be found in our released codes.

2.2 Loss Function

As aforementioned, human experts can make diagnosis more accurate by utilizing diverse imaging characteristics of various modalities. For instance, the enhancing tumor has a noticeable increase in T1 signal on post-contrast images relative to pre-contrast images. The surrounding non-enhancing flair hyper-intensity is better displayed in the FLAIR signal [9]. The pixel intensity difference between two modality images can serve as pseudo-labels for specific tissues and organs. Although they are insufficient for training a network due to the high noise, the offset between the two modality images between different patients always contains information about the same anatomy. Therefore, we hypothesize that the multi-modality image features should also have consistent and stable offsets and propose inter-modality contrastive loss and consistent loss to achieve this goal.

Inter-Modality Contrastive Learning. We first make multi-modality image features discriminable by using contrastive learning [11, 14]. It can learn an embedding space in which positive pairs stay close to each other while negative ones are far apart. In this case, the positive pair comprised features of the same modality images obtained by two encoders, and the negative pair comprised features of different modalities. Specifically, a global average pooling layer first transforms the features $f_{4,m}$ into $r_{4,m} \in R^{4,8C}$. Subsequently, two linear projectors map the encoder output features $f_{4,m}^{seq}$ and $f_{4,m}^{spa}$ into low-dimensional representations $r_m^{seq}, r_m^{spa} \in R^{4,512}$. Each projector consists of a linear layer followed by a ReLU activation layer. The L_{con} is defined as:

$$L_{con} = -log \frac{exp(sim(r_i^{spa}, r_i^{seq}))/t)}{\sum_j^m 1_{j \neq i}(sim(r_i^{spa}, r_j^{seq})/t)}, \tag{3}$$

where $t = 0.8$ is the temperature scale, 1 is the indicator function evaluating to 1 if $i \neq j$. We use cosine distance to measure the similarity between two features.

Inter-Modality Consistent Learning. The inter-modality consistent loss aims to enhance the anatomical information among different modalities. We achieve this by making the offsets between two modality image features consistent in two branches. We counted the pixel intensity difference maps of all two-modality combinations and used Shannon entropy to measure their information gain. The difference maps of the T_{1c} and T_2 contrasts has the highest entropy value of 3.64 and a smaller standard deviation of 0.48. Conversely, the T_1 and T_f contrasts exhibit the least entropy value of 3.50 and a larger standard deviation of 0.53. We calculate the L_{consis} between T_{1c} and T_2, T_1 and T_2, T_f and T_2. Specifically, we first apply a global average pooling layer to obtain flattened features. We then concatenate the paired modality features and use a multilayer perceptron (MLP) to derive the input features $v_{ij}^{seq}, v_{ij}^{spa} \in R^{512}$. For instance, $v_{01}^{seq} = MLP(Cat(Pooling(f_{4,0}^{seq}), Pooling(f_{4,1}^{seq})))$. The L_{consis} follows the same definition as L_{con} but sets the temperature t to 0.9. The positive pair is the features in two branches obtained by the same modality combination. The negative pair is the feature of different modality combinations.

Overall Loss Function. We use L_{rec} to learn the appearance of the anatomical structure in each modality. It is calculated by the Mean Square Error (MSE) in the pixel space between the original and the synthesized MRI images, given by

$$L_{rec} = \sum_{i=1}^{m} E[||x_m - \hat{x}_m||_2]. \tag{4}$$

Finally, our overall loss is:

$$L = \lambda_1 L_{rec} + \lambda_2 L_{con} + \lambda_3 L_{consis} \tag{5}$$

The $\lambda_1 = 1$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.5$ are the weights for each loss term.

Table 1. Quantitative comparison results of our method and competing methods on the BraTS dataset. ● means available real images, and ○ means imputed images. The best results are shown in bold. * means training without L_{consis}.

Modalities				PSNR ↑				SSIM ↑			
T_1	T_2	T_{1c}	T_f	Zhang	MMT	Ours*	Ours	Zhang	MMT	Ours*	Ours
●	○	○	○	**28.72**	26.77	26.65	28.02	**0.712**	0.542	0.637	0.648
○	●	○	○	**27.90**	26.76	26.82	27.14	**0.708**	0.534	0.636	0.651
○	○	●	○	**28.10**	26.21	26.04	27.43	**0.716**	0.553	0.602	0.614
○	○	○	●	**28.73**	27.39	27.44	28.67	**0.671**	0.510	0.612	0.621
○	○	●	●	27.73	27.78	27.74	**29.77**	0.791	0.794	0.865	**0.887**
○	●	●	○	27.79	27.74	27.60	**29.01**	0.786	0.757	0.813	**0.822**
●	●	○	○	27.73	27.87	27.66	**29.13**	0.761	0.765	0.821	**0.844**
○	●	○	●	27.50	27.56	27.65	**28.98**	0.772	0.745	0.808	**0.822**
●	○	●	○	27.69	27.77	27.92	**29.14**	0.738	0.696	0.787	**0.797**
●	●	●	○	28.99	28.88	29.86	**30.27**	0.821	0.841	0.887	**0.897**
○	●	●	●	28.77	28.87	30.75	**31.33**	0.833	0.861	0.913	**0.924**
●	●	○	●	30.03	30.09	30.03	**30.24**	0.828	0.847	0.883	**0.894**
●	○	●	●	29.18	29.12	29.19	**30.42**	0.844	0.853	0.897	**0.915**

3 Experiments

Dataset and Implementation Details. We build a 2D dataset sampled from the BraTS 2023 dataset Task 1. The BraTS dataset consists of 1251 annotated scans, each case contains four modalities. The segmentation performance is evaluated on WT, TC, and ET regions. We randomly selected 800, 200, and 250 cases for training, validation, and testing, respectively. We sampled 6 to 10 slices from each 3D volume at equal spacing along the coronal direction from all tumor-bearing axial slices. In this form, we get 6816, 1696, and 2240 paired 2D train, validation, and test dataset.

Our method was implemented in PyTorch 2.0 on a NVIDIA A100 40G GPU. The input size of each image modality is 256 × 256 pixels and batch size is set to 32. All images are non-zero normalized. The data augmentation includes random flip, intensity shift, and intensity scale with probability 0.5, 0.1 and 0.1. Our model was trained with Adam optimizer with an initial learning rate of $1e-4$ for 100 epochs.

Competing Methods and Evaluation Metrics. We compared our model with two missing modality imputation models and one segmentation model for modality missing. **MMT** [7] utilized modality encodings and modality queries to inject modality-specific information, which are learnable parameters for each modality. **Zhang et al.** [19] proposed a GAN-based unified model which can leverage both modality-invariant and modality-specific information and combine them with hard integration and soft integration to avoid information loss. **M^{3}FeCon** [18], abbreviated as M^3, treated missing modalities as masked modalities, and learned supplemental features similar to masked image modeling to form approximate modality-complete feature representations. The MMT and M^{3}FeCon were originally developed for 3D tasks and we re-implemented their

Table 2. Tumor segmentation evaluation on the BraTS dataset. ● means available real images, and ○ means imputed images. The best results are shown in bold. * means training without L_{consis}.

Modalities				WT				TC				ET			
T_1	T_2	T_{1c}	T_f	MMT	M^3	Ours*	Ours	MMT	M^3	Ours*	Ours	MMT	M^3	Ours*	Ours
●	○	○	○	79.26	78.41	79.08	**81.23**	65.24	64.96	67.32	**69.21**	53.85	52.20	54.96	**57.91**
○	●	○	○	89.11	89.23	88.24	**91.23**	64.49	64.50	64.68	**67.65**	52.51	48.29	53.33	**55.73**
○	○	●	○	80.49	81.28	81.88	**82.34**	72.37	72.00	74.68	**75.31**	59.28	61.21	59.21	**64.32**
○	○	○	●	88.95	87.87	88.56	**90.01**	71.16	69.17	70.15	**73.12**	51.08	49.39	53.96	**56.13**
○	○	●	●	88.19	88.01	88.14	**90.31**	77.56	77.60	80.06	**82.11**	71.43	71.26	74.34	**76.12**
○	●	●	○	84.72	83.82	89.88	**90.43**	80.29	79.63	81.31	**83.42**	72.86	73.19	75.20	**77.64**
●	●	○	○	79.89	80.78	82.20	**84.24**	70.71	71.68	75.59	**76.21**	65.05	64.40	68.22	**70.12**
○	●	○	●	80.08	79.52	83.86	**85.71**	69.59	70.47	74.10	**75.43**	67.56	68.70	70.97	**71.24**
●	○	●	○	86.72	86.59	88.08	**90.34**	82.34	83.09	83.50	**84.21**	74.65	75.77	77.43	**78.32**
●	●	●	○	88.59	89.24	88.20	**90.93**	86.07	**87.70**	86.52	86.93	83.45	**84.49**	82.09	83.78
○	●	●	●	91.48	92.43	90.96	**92.73**	86.40	86.84	86.53	**87.42**	80.91	80.98	82.22	**84.31**
●	●	○	●	87.99	**88.85**	88.01	88.31	79.44	79.88	79.27	**80.23**	58.97	58.53	62.11	**64.12**
●	○	●	●	92.48	89.10	91.14	**92.55**	87.22	86.84	87.16	**88.23**	81.92	82.31	84.86	**85.21**
Baseline				0.9380				0.9237				0.8692			

2D versions in all experiments. We use SSIM and PSNR as the synthetic image evaluation metrics and Dice score for segmentation evaluation.

3.1 Quantitative Evaluation

Table 1 presents the quantitative comparison results between our method and competing methods. We report the average results on generated missing modalities across all input-output combinations when one to three modalities are missing. In scenarios where only a single modality is accessible, Zhang's method, which employs a specialized GAN generator network, exhibits distinct advantages. Our method, however, attains superior PSNR and SSIM values in the majority of input scenarios. The advantage of our method in synthesis performance improves with more input modalities, indicating that our method can better utilize the inter-modality information in the inputs. Table 2 shows the tumor segmentation results to validate the utility of synthetic results. Our model improves the dice scores compared to previous state-of-the-art methods. We also conduct an ablation study on the contribution of proposed inter-modality consistency learning. The results in Table 1 and Table 2 indicate that incorporating consistency loss increases the performance demonstrating the effectiveness of the consistency learning.

Notably, within the context of our empirical experiments, it was observed that the fusion block plays an essential role in the early convergence of the sequence branch. This phenomenon was particularly evident when two to three out of four modalities were excluded. A plausible explanation is that, due to the relatively short length of the input sequence of the sequential branch, the inter-modality information becomes insufficient when more modalities are missing.

Thus, during the early convergence stage, we randomly mask only one modality and increase the number of masked modalities in a scheduled manner.

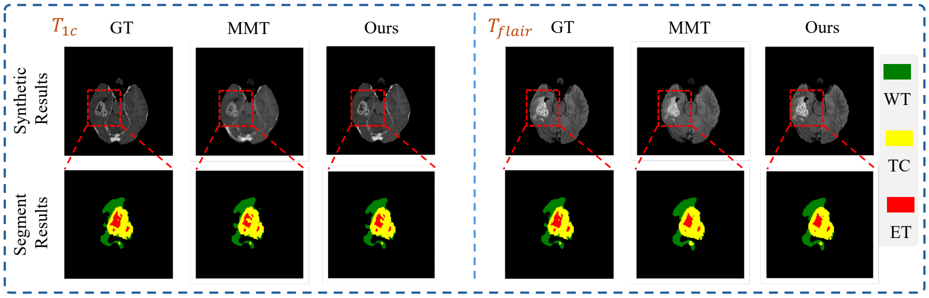

Fig. 2. Visual examples of synthetic T_{1c} and T_f modality images and corresponding segmentation results. The red boxes emphasize the tumor region.

3.2 Qualitative Evaluation

We observed the absence of T_{1c} modality and T_f results in particularly evident segmentation performance degradation in the ET and WT regions, respectively. Thus we visualize the synthetic T_{1c} and T_f images utilizing three other modalities and their segmentation results in Fig 2. Compared with MMT, our method produces better results with less blur and preserves more details.

4 Conclusion

In this paper, we propose a novel missing modality imputation model capable of generating missing modality images based on arbitrary available modalities. Our main contributions are as follows: an imputation backbone and an inter-modality contrastive and consistent learning strategy. Our method can generate high-quality images by leveraging the imaging characteristics of each modality and the shared anatomical information. Through quantitative and qualitative experiments, we have demonstrated that the generated images possess better fidelity. Moreover, when applied to the missing modality segmentation task, our method proves to be more robust than competing methods, thus validating the utility of the generated images.

One of the primary limitations of our approach is its subpar performance in scenarios where only a single modality is accessible. In future work, we intend to address this issue by maintaining a learnable dictionary during the training process. This dictionary can then be utilized for querying purposes when synthesizing images. Additionally, we will plan to extend our method to 3D tasks. The significant challenge lies in the contrastive loss we employed. This loss function generally requires large batch sizes to avert trivial solutions. However, due to memory constraints, a small batch size will lead to the failure of contrastive learning.

Acknowledgments. This study was funded by Jilin Provincial Scientific and Technological Development Program, China. Grant 20240305040YY.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)
2. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
4. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
6. Li, Y., Zhou, T., He, K., Zhou, Y., Shen, D.: Multi-scale transformer network with edge-aware pre-training for cross-modality mr image synthesis. IEEE Transactions on Medical Imaging **42**(11), 3395–3407 (2023)
7. Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. IEEE Transactions on Medical Imaging **42**(9), 2577–2591 (2023)
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
9. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
10. Novosad, P., Carano, R.A., Krishnan, A.P.: A task-conditional mixture-of-experts model for missing modality segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 34–43. Springer (2024)
11. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

13. Sharma, A., Hamarneh, G.: Missing mri pulse sequence synthesis using multi-modal generative adversarial network. IEEE transactions on medical imaging **39**(4), 1170–1183 (2019)
14. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
15. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
16. Yang, Q., Guo, X., Chen, Z., Woo, P.Y., Yuan, Y.: D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. IEEE Transactions on Medical Imaging **41**(10), 2953–2964 (2022)
17. Yuan, W., Wei, J., Wang, J., Ma, Q., Tasdizen, T.: Unified generative adversarial networks for multimodal segmentation from unpaired 3d medical images. Medical Image Analysis **64**, 101731 (2020)
18. Zeng, Z., Peng, Z., Yang, X., Shen, W.: Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–433. Springer (2024)
19. Zhang, Y., Peng, C., Wang, Q., Song, D., Li, K., Zhou, S.K.: Unified multi-modal image synthesis for missing modality imputation. IEEE Transactions on Medical Imaging (2024)
20. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal mr image synthesis. IEEE transactions on medical imaging **39**(9), 2772–2781 (2020)