DC-Seg: Disentangled Contrastive Learning for Brain Tumor Segmentation with Missing Modalities

Haitao Li, Ziyu Li, Yiheng Mao, Zhengyao Ding, and Zhengxing Huang (✉)

Zhejiang University
lihaitao@zju.edu.cn, zhengxinghuang@zju.edu.cn

Abstract. Accurate segmentation of brain images typically requires the integration of complementary information from multiple image modalities. However, clinical data for all modalities may not be available for every patient, creating a significant challenge. To address this, previous studies encode multiple modalities into a shared latent space. While somewhat effective, it remains suboptimal, as each modality contains distinct and valuable information. In this study, we propose DC-Seg (Disentangled Contrastive Learning for Segmentation), a new method that explicitly disentangles images into modality-invariant anatomical representation and modality-specific representation, by using anatomical contrastive learning and modality contrastive learning respectively. This solution improves the separation of anatomical and modality-specific features by considering the modality gaps, leading to more robust representations. Furthermore, we introduce a segmentation-based regularizer that enhances the model's robustness to missing modalities. Extensive experiments on the BraTS 2020 and a private white matter hyperintensity(WMH) segmentation dataset demonstrate that DC-Seg outperforms state-of-the-art methods in handling incomplete multimodal brain tumor segmentation tasks with varying missing modalities, while also demonstrate strong generalizability in WMH segmentation. The code is available at https://github.com/CuCl-2/DC-Seg.

Keywords: Brain Tumor Segmentation · Multi-modal · Missing Modality · Contrastive Learning · Disentangled Learning.

1 Introduction

Accurate brain image segmentation is crucial for assessing disease progression and developing effective treatments. Brain MRI, with modalities like T1, T2, T1ce, and FLAIR, provides varying sensitivity to lesion regions depending on imaging parameters and protocols [5]. Joint learning across these multimodal images improves segmentation accuracy compared to single-modality approaches. Common methods involve concatenating images from different modalities [9, 26, 2] or integrating features in high-dimensional spaces [6, 19, 24]. However, missing modalities due to protocol variations or patient factors pose a challenge.

Significant efforts have been made to address the challenges posed by missing modalities in practical scenarios, with existing solutions falling into three main categories. The first approach synthesizes missing modalities to complete the test set [20, 18, 13]. This involves training a generative model to generate missing modalities, but it often requires additional training and struggles when only one modality is available during inference. The second approach trains a dedicated model for each specific missing-modal scenario. Methods like [8, 4, 21, 1] distill knowledge from a multimodal teacher network to monomodal students at the image and pixel levels. Considering the varying sensitivities of lesion regions across modalities, GSS [16] selects a group leader for distillation. While these methods perform well when multiple modalities are missing, they incur high computational and memory costs, requiring $2^N - 1$ models for N modalities. The third approach attempts to handle all missing-modal situations with a single unified model, embedding all modalities into a shared latent space, followed by feature fusion for segmentation [7, 3]. RFNet [5] uses a region-aware fusion module to adaptively combine features from available modalities on different regions, while mmFormer [25] leverages Transformer for long-range dependencies, and M^{3}AE [11] employs multimodal autoencoders to reduce model complexity by creating a unified latent representation.

While valuable, these studies often overlook modality gaps, failing to learn invariant feature representations across modalities, which impairs performance in missing-modality scenarios. To address this, some approaches [3, 1, 22] decompose images into modality-invariant and modality-specific components, using invariant representations for segmentation. For instance, SMU-Net [1] posits that deeper network layers capture content representations, while shallower layers preserve style representations. Similarly, RobustSeg [3] decouples content and appearance codes by reconstructing images. D2Net [22] learns modality-specific codes through contrastive learning applied to different MRI slices.

In this study, we introduce bidirectional contrastive learning, complementing the traditional reconstruction task to achieve effective decoupling. Unlike previous models, our approach applies both anatomical and modality contrastive learning at the 3D MRI image level. This allows us to learn not only modality representations but also modality-invariant anatomical representations which are crucial for accurate segmentation. By performing contrastive learning on the full 3D image, we achieve more comprehensive feature extraction. Specifically, anatomical contrastive learning pulls features from the same individual across modalities closer, while pushing features from different individuals apart. Similarly, modality contrastive learning pulls features from the same modality across different individuals closer while pushing features from different modalities apart. Additionally, a segmentation-based regularizer is incorporated to further enhance the model's robustness to incomplete modalities.

We validate our method on the BRATS [14] dataset for multimodal brain tumor segmentation, achieving competitive performance in full-modality scenarios and superior robustness in missing-modality settings. Additionally, we demonstrate its generalizability on a private WMH segmentation dataset.

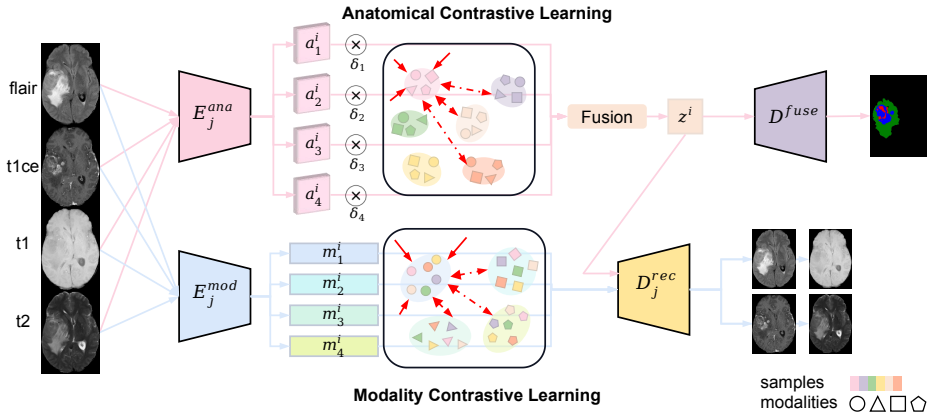

Fig. 1: Overview of DC-Seg, which disentangles images from different modalities into anatomical and modality representations using bidirectional contrastive learning, and fuses modality-invariant anatomical representations for the downstream tumor segmentation task. For clarity in the figure, D^{sep} is omitted.

2 Method

An overview of our proposed DC-Seg is shown in Fig. 1. First, we decouple the multimodal inputs into modality-specific and modality-invariant anatomical representations using both the traditional reconstruction task and our proposed novel bidirectional contrastive learning approach. Next, we fuse the anatomical representations from different modalities for tumor segmentation. Additionally, a segmentation-based regularizer is introduced to prevent the model from becoming highly dependent on discriminative modalities (e.g., T1ce, FLAIR) for brain tumor recognition, which could lead to significant degradation in performance when these discriminative modalities are missing. The detailed learning process and network architecture are described below.

2.1 Bidirectional Contrastive Learning

In this part, we describe how images are disentangled into anatomical and modality-specific representations. Traditional methods [10, 15, 3] achieve this by reconstructing images from fused anatomical representation and modality-specific representations across all modalities. However, these methods only ensure that the fused anatomical representation is well-learned and do not guarantee that the anatomical representation for each modality is modality-invariant, which can lead to performance degradation when modalities are missing. To address this issue, we introduce bidirectional contrastive learning to ensure that anatomical representations across all modalities are effectively learned and aligned.

Here's how it works, given a batch of multimodal images $\{x_j^i\}$, where $i \in \{1, \ldots, N\}$ indexes the samples and $j \in \{1, \ldots, M\}$ indexes the modalities, with

$M = 4$ in our brain tumor segmentation task, each modality x_j^i from sample x^i is passed through its respective anatomical encoder E_j^{ana} and modality encoder E_j^{mod} to obtain corresponding disentangled anatomical representation $a_j^i = E_j^{\text{ana}}(x_j^i)$ and modality representation $m_j^i = E_j^{\text{mod}}(x_j^i)$. For the modality representation, we follow the common practice in [10], representing it as an 8-bit vector $m_j^i \in \mathbb{R}^C$, while the anatomical representation is encoded as multichannel 3D feature maps $a_j^i \in \mathbb{R}^{C\times d\times d\times d}$. Intuitively, we aim to align different modalities of the same sample to share a common anatomical representation, which is essential for downstream tasks like brain tumor segmentation. For modality-specific representations, we ensure that images from the same modality are closely aligned. To achieve these objectives, we introduce bidirectional contrastive learning.

In anatomical contrastive learning, given an anchor image x_j^i, the positive samples are images from the same subject, denoted as $x_{j'}^{i'}$, $i' = i$ while the negative samples are from different subjects ($i' \neq i$). Since each anchor image has multiple positive samples, we do not use the softmax-based contrastive loss as in CLIP [17]. Instead, we employ a sigmoid-based loss similar to [23]. Formally, the anatomical contrastive loss L_{ana} is defined as:

$$\mathcal{L}_{\text{ana}} = -\frac{1}{(N\cdot M)^2}\sum_{i,j,i',j'}\log\frac{1}{1+e^{f(i,i')\cdot\left(-t\cdot\text{SSIM}(a_j^i,a_{j'}^{i'})\right)}} \tag{1}$$

$$f(i,i') = \begin{cases} 1, & \text{if } i = i', \\ -1, & \text{if } i \neq i' \end{cases} \tag{2}$$

where a batch containing $N \times M$ images results in $(N \times M)^2$ pairs. $f(i, i')$ is used to determine whether two images belong to the same sample. The temperature scaling factor t controls the sharpness of the distribution, influencing the model's sensitivity to positive and negative pairs. The $\text{SSIM}(a_j^i, a_{j'}^{i'})$ defined below represents the channel-wise mean of structural similarity between the feature maps $a_{j,c}^i$ and $a_{j',c}^{i'}$ for each channel. The constants C_1 and C_2 are small values introduced to prevent division by zero and stabilize the computation.

$$\text{SSIM}(a_j^i, a_{j'}^{i'}) = \frac{1}{C}\sum_{c=1}^{C}\frac{(2\mu_{a_{j,c}^i}\mu_{a_{j',c}^{i'}} + C_1)(2\sigma_{a_{j,c}^i,a_{j',c}^{i'},} + C_2)}{(\mu_{a_{j,c}^i}^2 + \mu_{a_{j',c}^{i'}}^2 + C_1)(\sigma_{a_{j,c}^i}^2 + \sigma_{a_{j',c}^{i'}}^2 + C_2)} \tag{3}$$

Similar to the anatomical contrastive loss in Eq. 1, the modality contrastive loss $\mathcal{L}_{\text{mod}}$ is defined as Eq. 4, with a key distinction: images from the same modality are treated as positive pairs (i.e., $j = j'$), while images from different modalities are treated as negative pairs. Since the modality representation $m_{i,j}$ is an 8-bit vector, cosine similarity is used instead of SSIM. The similarity is defined as: $\text{sim}(m_j^i, m_{j'}^{i'}) = \frac{m_j^i \cdot m_{j'}^{i'}}{\|m_j^i\|_2\|m_{j'}^{i'}\|_2}$.

$$\mathcal{L}_{\text{mod}} = -\frac{1}{(N\cdot M)^2}\sum_{i,j,i',j'}\log\frac{1}{1+e^{f(j,j')\cdot\left(-t\cdot\text{sim}(m_j^i,m_{j'}^{i'})\right)}} \tag{4}$$

In addition to the bidirectional contrastive learning discussed above, we adhere to the assumption that, for successful disentanglement, the obtained anatomical representation should be re-renderable into the original image when paired with the modality representation of any given modality [10]. Specifically, we fuse the anatomical representations from different modalities to obtain z^i following [5], and then reconstruct the image using a set of modality-specific decoders, $\{D_j^{\text{rec}}\}$, given z^i and the modality representation m_j^i. The loss function is defined below, where we use the L1-norm to prevent image blurring. A Bernoulli indicator δ_i is employed to enhance the robustness of the content representation z to missing data, with modality dropout applied in the latent space by randomly setting δ_i to 0.

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N}\sum_{j=1}^{M} \|D_j^{\text{rec}}(z^i, m_j^i) - x_j^i\|_1, \quad \text{where} \quad z^i = \mathcal{F}(\delta_1 a_1^i, \delta_2 a_2^i, \ldots, \delta_M a_M^i), \tag{5}$$

The final disentanglement loss is defined as follows.

$$\mathcal{L}_{\text{disentangle}} = \mathcal{L}_{\text{ana}} + \mathcal{L}_{\text{mod}} + \mathcal{L}_{\text{rec}} \tag{6}$$

2.2 Learning Process

Due to the high sensitivity of certain discriminative modalities (e.g., T1ce, FLAIR) to specific tumor regions, the model tends to depend on these modalities for segmentation, resulting in significant performance degradation when they are unavailable. Therefore, it is critical to encourage the model to segment based on all modalities. To achieve this, we introduce a segmentation-based regularizer like [5, 25]. Specifically, we use a weight-shared decoder D^{sep} to segment based on every single modality separately. The corresponding weighted cross-entropy loss and Dice loss are used as regularization terms, expressed as:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{N}\sum_{j=1}^{M} \left(\mathcal{L}_{\text{WCE}}(D^{\text{sep}}(a_j^i), y^i) + \mathcal{L}_{\text{DL}}(D^{\text{sep}}(a_j^i), y^i)\right), \tag{7}$$

As illustrated in Fig. 1, the fused anatomical feature z^i is used to predict the final segmentation mask through D^{fuse}. The weighted cross-entropy loss and Dice loss are employed to align the predictions with the corresponding ground-truth segmentation maps, as expressed below:

$$\mathcal{L}_{\text{seg}} = \sum_{i=1}^{N} \left(\mathcal{L}_{\text{WCE}}(D^{\text{fuse}}(z^i), y^i) + \mathcal{L}_{\text{DL}}(D^{\text{fuse}}(z^i), y^i)\right), \tag{8}$$

Therefore, the overall loss of our DC-Seg is defined below, with α as a hyperparameter for the tradeoff.

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{reg}} + \alpha\mathcal{L}_{\text{disentangle}} \tag{9}$$

Table 1: Results of the proposed method and state-of-the-art unified models on BraTS 2020 dataset. Dice similarity coefficient is employed for evaluation with every combination of modality settings. • and ◦ denote available and missing modalities, respectively.

Modalities				Complete					Core					Enhancing				
F	T1	T1c	T2	RobustSeg	RFNet	mmFormer	M^{3}AE	**DC-Seg**	RobustSeg	RFNet	mmFormer	M^{3}AE	**DC-Seg**	RobustSeg	RFNet	mmFormer	M^{3}AE	**DC-Seg**
◦	◦	◦	•	82.20	86.05	85.51	86.10	**86.72**	61.88	71.02	63.36	**71.80**	70.88	36.46	46.29	**49.09**	47.10	47.76
◦	◦	•	◦	71.39	76.77	78.04	78.90	**79.54**	76.68	81.51	81.51	83.60	**84.62**	67.91	74.85	78.30	73.60	**78.90**
◦	•	◦	◦	71.41	77.16	76.24	**79.00**	78.47	54.30	66.02	63.23	**69.40**	66.63	28.99	37.30	37.62	40.40	**42.19**
•	◦	◦	◦	82.87	87.32	86.54	**88.00**	87.80	60.72	69.19	64.60	68.70	**71.27**	34.68	38.15	36.68	40.20	**41.66**
◦	◦	•	•	85.97	87.74	87.52	87.10	**88.17**	82.44	83.45	82.69	85.60	**86.34**	71.42	75.93	77.20	76.00	**80.43**
◦	•	•	◦	76.84	81.12	80.70	80.10	**82.22**	80.28	83.40	82.81	83.80	**85.18**	70.11	78.01	**81.71**	75.30	79.25
•	•	◦	◦	88.10	89.73	88.76	89.60	**90.01**	68.18	73.07	71.76	72.80	**74.50**	39.67	40.98	42.98	43.70	**46.90**
◦	•	◦	•	85.53	87.73	86.94	87.30	**88.09**	66.46	**73.13**	67.76	72.90	73.09	39.92	45.65	49.12	48.70	**50.19**
•	◦	◦	•	88.09	89.87	89.49	90.10	**90.32**	68.20	74.14	70.34	74.30	**75.11**	42.19	49.32	49.06	47.10	**51.32**
•	◦	•	◦	87.33	89.89	89.31	89.50	**89.99**	81.85	84.65	83.79	85.50	**85.90**	70.78	76.67	79.44	75.90	**80.28**
•	•	•	◦	88.87	**90.69**	89.79	89.60	90.65	82.76	85.07	84.44	85.60	**86.29**	71.77	76.81	80.65	76.30	**81.41**
•	•	◦	•	89.24	90.60	89.83	90.20	**90.77**	70.46	75.19	72.42	74.40	**75.53**	43.90	49.92	50.08	48.20	**52.05**
•	◦	•	•	88.68	**90.68**	90.49	90.50	90.62	81.89	84.97	83.94	85.80	**86.21**	71.17	77.12	78.73	77.40	**79.42**
◦	•	•	•	86.63	88.25	87.64	87.40	**88.73**	82.85	83.47	83.66	85.80	**86.49**	71.87	76.99	77.34	78.00	**81.66**
•	•	•	•	89.47	**91.11**	90.54	90.40	90.95	82.87	85.21	84.61	86.20	**86.46**	71.52	78.00	79.92	77.50	**81.52**
Average				84.17	86.98	86.49	86.90	**87.54**	73.45	78.23	76.06	79.10	**79.63**	55.49	61.47	63.19	61.70	**65.00**

3 Experiments and Results

Datasets and Implementation. The experiments are conducted using the BraTS 2020 dataset [14] and a private white matter hyperintensity segmentation dataset(SAHZU-WMH) from The Second Affiliated Hospital, Zhejiang University School of Medicine.

The BraTS 2020 dataset includes 369 multi-contrast MRI scans across four modalities (T1, T1c, T2, FLAIR), with tumor subregions: whole tumor, tumor core, and enhancing tumor. All volumes are skull-stripped, co-registered to a common anatomical template, resampled to 1mm^3 isotropic resolution, and normalized to zero mean and unit variance within the brain tissue. Patches of size $112 \times 112 \times 112$ are randomly cropped and used as input to the network during training.

The SAHZU-WMH dataset includes 41 patients with cognitive impairment, each with two follow-up MRI scans (average interval of 473 days), totaling 80 multi-modal scans (FLAIR and T1) after excluding one missing follow-up and one incorrect annotation. White matter hyperintensity regions are annotated. Preprocessing follows the same steps as BraTS 2020, with patches of size $128 \times 128 \times 128$. The dataset is split into 60 scans from 31 individuals for training and 20 scans from 10 individuals for testing, ensuring no overlap between subjects in the training and test sets.

Random flips, cropping, and intensity shifts are applied for data augmentation. The network is trained using the Adam optimizer with an initial learning rate of 0.0002 for 500 epochs with batch size 2. The hyperparameters are set as: $\alpha = 0.4$.

Performance of Incomplete Multimodal Segmentation. We evaluate the robustness of our method for incomplete multimodal segmentation. The absence of a modality is simulated by setting δ_i to zero. We compare our method against state-of-the-art approaches, namely RobustSeg [3], RFNet [5], mmFormer

[25], M^{3}AE [11] and GSS[16]. For a fair comparison, we use the same data split as in [5] and directly reference the results. As shown in Table 1, our method significantly outperforms the state-of-the-art methods in the segmentation of all three tumor parts across most of the 15 possible modality combinations. Moreover, our approach surpasses the large pre-trained M^{3}AE model [11] (Complete: 86.9, Core: 79.1, Enhancing: 61.7) and achieves comparable performance to the dedicated method GSS [16] (Complete: 87.33, Core: 79.38, Enhancing: 65.54), which requires 15 models for all modalities combination. Figure 2 illustrates that our method effectively segments brain tumors across various missing modality scenarios. Additionally, we also tested the medical foundation model MedSAM [12] for segmenting the complete tumor. However, even when provided with a bounding box as a prompt, MedSAM fails to clearly delineate the tumor's contour, further demonstrating the continued significance of our approach, even in the era of prevalent foundation models.

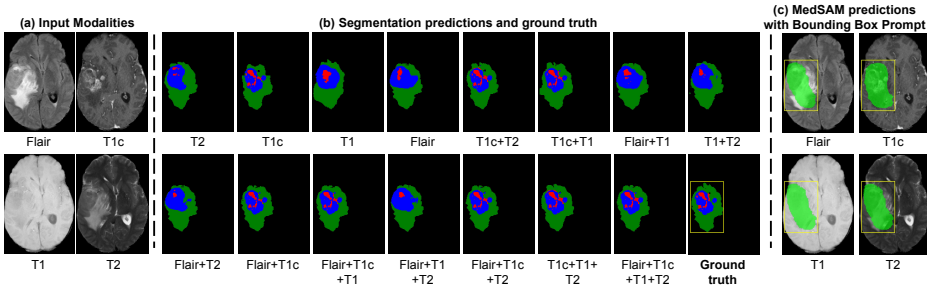

Fig. 2: Visualization of the input modalities, our predicted segmentation maps, and MedSAM prediction with bounding box prompt.

Disentanglement Visualization Figure 3 visualizes the anatomical and modality representations on the unseen BRATS test set. After anatomical contrastive learning, modality-invariant anatomical representations are effectively aligned for each modality, improving the model's robustness to missing modalities. Additionally, modality-specific representations are successfully learned.

Ablation Study. We investigate the effectiveness of anatomical contrastive learning, modality contrastive learning, the reconstruction task, and the regularizer as key components of our method. To assess the contribution of each component, we evaluate the performance of DC-Seg with each component excluded. In Table 2, we compare the performance of these variants against the full DC-Seg model, measured by the Dice Similarity Coefficient (DSC), averaged over the 15 possible combinations of input modalities. The results show that each component contributes to performance improvement across all tumor subregions.

Performance of WMH Segmentation. We evaluate the performance of our method for WMH segmentation across various modality combinations. The

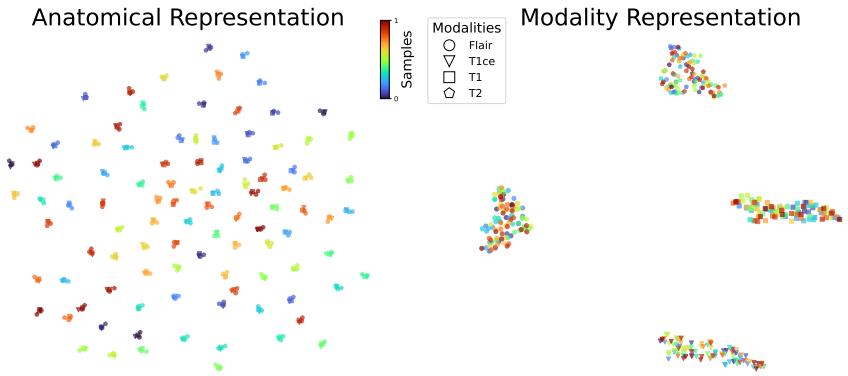

Fig. 3: Visualization of anatomical and modality representations on the unseen test set of BRATS.

Table 2: Ablation study. *Ana and Mod: Anatomical and modality contrastive learning, Rec: reconstruction, Reg: Regularizer*

Ana	Mod	Rec	Reg	Complete	Core	Enhancing
✓	✓	✓	✓	87.54	79.63	65.00
×	×	✓	✓	85.64	76.26	61.83
×	✓	✓	✓	85.81	77.15	62.37
✓	×	✓	✓	86.31	77.53	63.41
✓	✓	×	✓	86.58	77.54	64.48
✓	✓	✓	×	86.38	76.58	63.40

Table 3: Results of WMH Segmentation. Dice similarity coefficient is employed for evaluation

Modalities		**Dice scores**			
Flair	**T1**	**RobustSeg**	**RFNet**	**mmFormer**	**Ours**
○	•	55.56	58.26	53.91	**60.42**
•	○	81.69	82.35	77.29	**83.03**
•	•	82.24	82.66	77.98	**82.84**
Average		73.16	74.42	69.73	**75.43**

results, shown in Table 3, highlight the effectiveness of our approach in comparison to state-of-the-art methods. Our method consistently outperforms all other methods across all modality combinations. Notably, in the case where only the T1 modality is available—where white matter hyperintensities are particularly challenging to discern—our approach still achieves the highest performance. These results emphasize the superior performance of our approach in WMH segmentation besides tumor segmentation, demonstrating its generalizability.

4 Conclusion

We propose DC-Seg, a novel multimodal segmentation framework that jointly uses anatomical contrastive learning and modality contrastive learning to decompose images into modality-invariant anatomical representations and modality-specific representations. We demonstrate the superiority of DC-Seg through extensive experiments on both the BraTS 2020 brain tumor dataset and a private white matter hyperintensity segmentation dataset, achieving state-of-the-art results in full-modality scenarios and outperforming existing methods in missing-modality conditions.

Acknowledgments. This study was supported by the Technical Innovation Key Project of Zhejiang Province (2024C03023).

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

References

1. Azad, R., Khosravi, N., Merhof, D.: Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In: International Conference on Medical Imaging with Deep Learning. pp. 48–62. PMLR (2022)
2. Chen, C., Liu, X., Ding, M., Zheng, J., Li, J.: 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 184–192. Springer (2019)
3. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 447–456. Springer (2019)
4. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging **41**(3), 621–632 (2021)
5. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)
6. Fidon, L., Li, W., Garcia-Peraza-Herrera, L.C., Ekanayake, J., Kitchen, N., Ourselin, S., Vercauteren, T.: Scalable multimodal convolutional networks for brain tumour segmentation. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 285–293. Springer (2017)
7. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 469–477. Springer (2016)
8. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 772–781. Springer (2020)
9. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)
10. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)

11. Liu, H., Wei, D., Lu, D., Sun, J., Wang, L., Zheng, Y.: M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1657–1665 (2023)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Meng, X., Sun, K., Xu, J., He, X., Shen, D.: Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. IEEE Transactions on Medical Imaging (2024)
14. 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)
15. Ouyang, J., Adeli, E., Pohl, K.M., Zhao, Q., Zaharchuk, G.: Representation disentanglement for multi-modal brain mri analysis. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. pp. 321–333. Springer (2021)
16. Qiu, Y., Chen, D., Yao, H., Xu, Y., Wang, Z.: Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21317–21326 (2023)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
18. Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., et al.: Multi-domain image completion for random missing input data. IEEE transactions on medical imaging **40**(4), 1113–1122 (2020)
19. Tseng, K.L., Lin, Y.L., Hsu, W., Huang, C.Y.: Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6393–6400 (2017)
20. Van Tulder, G., de Bruijne, M.: Why does synthesized data improve multi-sequence classification? In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 531–538. Springer (2015)
21. Wang, Y., Zhang, Y., Liu, Y., Lin, Z., Tian, J., Zhong, C., Shi, Z., Fan, J., He, Z.: Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 410–420. Springer (2021)
22. 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)
23. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
24. Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., Wang, Y., Yu, Y.: Exploring task structure for brain tumor segmentation from multi-modality mr images. IEEE Transactions on Image Processing **29**, 9032–9043 (2020)

25. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2022)
26. Zhou, C., Ding, C., Lu, Z., Wang, X., Tao, D.: One-pass multi-task convolutional neural networks for efficient brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11. pp. 637–645. Springer (2018)