This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Multi-modal Vision Pre-training for Medical Image Analysis

Shaohao Rui[1,2*], Lingzhi Chen[2*], Zhenyu Tang[1,2], Lilong Wang[2], Mianxin Liu[2], Shaoting Zhang[2], Xiaosong Wang[2✉]
[1]Shanghai Jiao Tong University, Shanghai, China [2]Shanghai AI Laboratory, Shanghai, China
{ruishaohao, tang_zhenyu}@sjtu.edu.cn
{chenlingzhi, wanglilong, liumianxin, zhangshaoting, wangxiaosong}@pjlab.org.cn
https://github.com/openmedlab/BrainMVP

## Abstract

*Self-supervised learning has greatly facilitated medical image analysis by suppressing the training data requirement for real-world applications. Current paradigms predominantly rely on self-supervision within uni-modal image data, thereby neglecting the inter-modal correlations essential for effective learning of cross-modal image representations. This limitation is particularly significant for naturally grouped multi-modal data, e.g., multi-parametric MRI scans for a patient undergoing various functional imaging protocols in the same study. To bridge this gap, we conduct a novel multi-modal image pre-training with three proxy tasks to facilitate the learning of cross-modality representations and correlations using multi-modal brain MRI scans (over 2.4 million images in 16,022 scans of 3,755 patients), i.e., cross-modal image reconstruction, modality-aware contrastive learning, and modality template distillation. To demonstrate the generalizability of our pre-trained model, we conduct extensive experiments on various benchmarks with ten downstream tasks. The superior performance of our method is reported in comparison to state-of-the-art pre-training methods, with Dice Score improvement of 0.28%-14.47% across six segmentation benchmarks and a consistent accuracy boost of 0.65%-18.07% in four individual image classification tasks.*

## 1. Introduction

Medical image analysis is greatly enhanced via self-supervised learning for its capability of extracting distinctive image representation and surprisingly robust generalization performance across various downstream applications. However, current self-supervised learning methods in medical imaging are still confined to pre-training on uni-modal image data, e.g., computed tomography (CT) imaging [16, 21, 56], magnetic resonance imaging (MRI) [33, 46, 48, 65], and X-rays [5, 17, 32, 47, 61], or mixed image data (with each modality processed separately) [24, 53, 57]. These methods mainly focus on instance-level discrimination [16, 27, 39, 53, 56] or image reconstruction [33, 46, 48, 65] proxy tasks, failing to effectively model the relationships across modalities. In clinical practice, as shown in Fig. 1(a), groups of multi-modal data are naturally acquired as different imaging protocols are set to capture complementary pathological features for the same patient in one examing study. In other words, these multi-modal data acquired on the same patient exhibit strong correspondences. For example, multi-parametric MRI (mpMRI) data combining diverse modalities helps to comprehensively depict the structural and pathological features of the brain [44], which substantially enhances diagnostic accuracy and thorough-

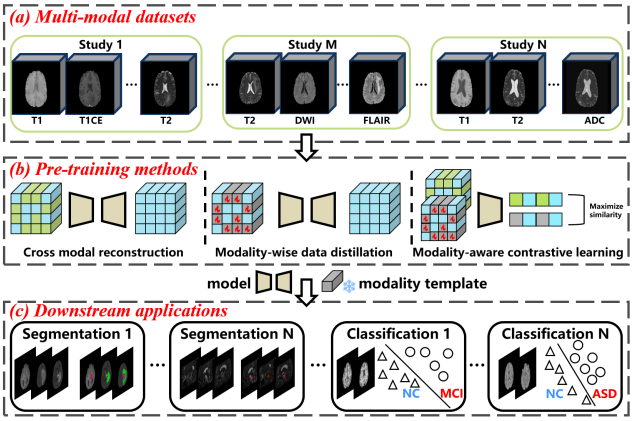



Figure 1. (a) There are naturally grouped multi-modal data, e.g., multi-parametric MRI scans in the real world. (b) We propose three proxy tasks to facilitate the learning of cross-modality representations and correlations. Blue cubes represent one modality in a study, green cubes represent another modality within the same study, and gray cubes with flame symbols represent learnable modality templates. (c) We apply the pre-trained model and distilled modality templates for downstream tasks.

∗ Equal contribution
✉ Corresponding author (wangxiaosong@pjlab.org.cn)

ness [43]. Therefore, it is crucial to develop novel self-supervised learning frameworks for such grouped multi-modal image data, tailored to the downstream applications in the aforementioned scenarios.

In the real world, another issue of model training and inference with multi-modal data is the presence of missing modalities. Obtaining a comprehensive set of modalities in mpMRI scans can be challenging due to the complexity associated with acquisition protocol and limitations in equipment capabilities. This often leads to mismatched modality data across datasets, especially when the scale of the data amount increases dramatically. Current approaches to deal with missing modalities primarily focus on particular downstream tasks, e.g., brain tumor segmentation in BraTS [2, 3], and have not undergone extensive investigation in large-scale cross-modal pre-training and its downstream applications [15, 28, 42, 50, 63].

In this paper, we introduce BrainMVP, a novel self-supervised learning framework designed for multi-modal MRI data, as illustrated in Fig. 1(b). The main goal of BrainMVP is to create generalizable cross-modal representations while effectively tackling the challenge of missing modalities during pre-training. Thus, we also compile a dataset of 16,022 publicly available brain mpMRI scans, sourced from a range of multi-center and multi-device contributions, to demonstrate our pre-training initiatives.

For the issue of insufficient scalability resulting from mismatched or missing modalities, we propose using uni-modal MRI inputs instead of fixed modality numbers [33] during pre-training. This allows for the inclusion of arbitrary numbers of modalities in the pre-training, significantly expanding the magnitude of available pre-training data. Moreover, we propose cross-modal reconstruction via masked image modeling. A key aspect of this design is the observation that different MRI modalities for the same patient often exhibit significant similarity in anatomy. By employing cross-modal reconstruction, we encourage the model to learn the disentanglement among modalities.

Towards a more generalizable pre-training model for downstream tasks, we also extract condensed structural representations of different modalities using modality-wise data distillation. Our approach is inspired by the technique of dataset distillation, which involves learning a small synthetic dataset. The performance achieved by the model training on this synthetic dataset can rival that achieved on the original large-scale datasets [52, 60, 64]. The learned synthetic dataset indeed encapsulates dense representations of the original dataset. In a similar idea, we optimize a set of learnable modality templates tailored for each individual modality. Intuitively, the distilled modality templates retain shared structural and statistical information about a specific modality while avoiding privacy leakage concerns associated with individual patients. Therefore, the distilled modality templates can serve as a linkage of data between pre-training and downstream tasks, i.e., as a form of information to carry and adapt between the data domains in downstream applications.

In summary, our contributions are three-fold:

- To the best of our knowledge, BrainMVP is the first multi-modal vision pre-training paradigm that aligns the features across modalities, targeting distinctive modality-aware representations. We also collect a dataset of 16,022 mpMRI scans (3,755 patients, over 2.4 million images) to facilitate the pre-training, covering a wide range of brain MRI scans in both diseased and healthy populations.
- We design two novel proxy tasks for cross-modal representation learning, i.e., cross-modal reconstruction and cross-modal contrastive learning. To improve the generalization for downstream tasks, we also introduce the third modality-wise data distillation task to extract compact templates for each modality, benefiting both the pre-training and downstream tasks.
- We demonstrate the superior performance and the enhanced generalizability of our BrainMVP pre-trained models on ten public segmentation and classification benchmarks compared to state-of-the-art methods.

## 2. Related Work

Given the typically limited datasets available for specific medical tasks, pre-training on large-scale unlabeled data to extract highly generalizable representations is emerging as a new paradigm. Existing SSL methods in medical imaging can be roughly divided into two categories: uni-modal SSL and multi-modal SSL (with mixed modality data). While there have been numerous advancements in multi-modal learning involving paired text knowledge injection [9, 49], we concentrate on representation learning within medical imaging in this paper.

**SSL using uni-modal data**: Due to the convenience of data collection and storage, many self-supervised learning methods based on uni-modal imaging have emerged. Typical uni-modal SSL researches include computed tomography (CT) imaging[16, 21, 56], magnetic resonance imaging (MRI) [33, 46, 48, 65], and X-rays [5, 17, 32, 47, 61]. While impressive results have been achieved in specific uni-modal tasks, models pre-trained on uni-modal data often excel only in that specific modality and lack strong generalization capabilities. For example, models pre-trained on natural images struggle to generalize to medical imaging scenarios, and models trained on CT images find it challenging to generalize to MR images.

**SSL using mixed modality data**: It has been validated that multi-modal data from different imaging sources can be unified through shared encoders in a self-supervised learning manner and also play a complementary role in promoting the representation learning of specific modali-

ties [33, 45, 48, 54]. Composing CT, X-ray, and MR images, PCRLv2 [62] addresses the issue of local information loss in medical images within the contrastive learning SSL paradigm by suggesting pixel recovery and feature alignment at various scales for diverse enhancement samples. Additionally, PCRLv2 [62] recommends implementing SSL without using skip connections to avoid shortcut solutions in pixel restoration. Using CT and X-ray images, VoCo [53] leverages the contextual position priors to learn consistent semantic representations in pre-training and performs exceptionally well in medical images where the relative positions are relatively fixed. Although the methods above involve joint training on multi-modal data, different data sources often pose a bottleneck to the model's cross-modal understanding. [44] introduces a multi-modal puzzle task designed to enhance representation learning from various image modalities and applies modal transformation based on a generative network while solely acting as a data augmentation strategy. Our method, instead, employs a simple yet effective strategy of cross-modal reconstruction to learn cross-modal representations, incorporating modal complementary properties into the pre-training proxy task.

**Dataset distillation for knowledge compression**: Dataset distillation is first proposed to distill a core set in which the learned model can achieve a performance comparable to that of the whole dataset [52]. In this way, computational burden and data storage costs can be significantly reduced [60]. Existing dataset distillation methods can mainly be categorized into three types: parameter matching [7, 12, 25, 30, 58, 60], distribution matching [51, 59] and performance matching [13, 34, 38]. The novel modality data distillation method presented in this paper is inspired by performance matching methods, where distilled templates are learned via reconstructing the real modality image along the pre-training trajectories. This approach first learns patient-agnostic structural representations within modalities and then integrates patient-specific modality information during downstream tasks to bridge the domain gap and enhance model generalization.

## 3. Methods

As shown in Fig. 2, BrainMVP comprises three key modules: cross-modal reconstruction, modality-wise data distillation, and modality-aware contrastive learning. Three modules are detailed in the following sections.

### 3.1. Cross-Modal Reconstruction

**Problem Setting**: Given an unlabeled dataset $\mathcal{D} = \{X_{im} \in \mathbb{R}^{D\times H\times W} | m \in \{1,\dots,M_i\}, i \in \{1,\dots,N\}\}$, where $M_i$ denotes the number of modalities in the $i$-th sample and $N$ represents total number of samples. Masked image modeling (MIM) first masks with noise or discards (denoted as $\Phi(\cdot)$ ) a large portion of $X_{im}$ to obtain a masked input $\Phi(X_{im})$, and then reconstructs the original image from it to learn efficient representations. Specifically, let the model be $\mathcal{F}(\cdot) = \mathcal{F}_{dec} \circ \mathcal{F}_{enc}(\cdot)$, where $\mathcal{F}_{\text{enc}}(\cdot)$ and $\mathcal{F}_{\text{dec}}(\cdot)$ are the encoder and decoder respectively, MIM minimizes the following reconstruction loss:

$$\mathcal{L}_{rec} = ||\mathcal{F}_{\text{dec}}(\mathcal{F}_{\text{enc}}(\Phi(X_{im})) - X_{im}||_2 \,. \quad (1)$$

The core idea of our proposed reconstruction proxy tasks, which are elaborated in Sections 3.1 and 3.2, is to obtain meaningful representations via exploiting different forms of $\Phi(\cdot)$ function.

**Pixel-level cross-modal masking.** Given a uni-modal input volume $X_{im}$ sampled from an mpMRI case (with $M_i$ modalities), cross-modal masking aims to mask out a large region of $X_{im}$ and replace with another modality image $X_{in}$ (also sampled from $X_i$, $n \neq m$). Specifically, we first randomly mask a region of size $r \times r \times r$ in $X_{im}$, where $r$ denotes the size of each dimension of 3D volumes. Then, we fill in the masked region with a patch cropped with the same location and size on another modality of the sampled case. Finally, we repeat the above masking-filling operation until the proportion of masked pixels over the total input volume ($X_{im}$) pixels arrives $p^*$. More details of the masking algorithm can be found in the supplementary materials.

**Cross-modal reconstruction.** Let our proposed cross-modal masking strategy be $\Phi_{modal}$. Given that the masking operation masks a large portion of the image, the resulting masked input volume $\Phi_{modal}(X_{im}, X_{in})$ will contain information predominantly from $X_{in}$. The extracted representation $\mathcal{F}_{enc}(\Phi_{modal}(X_{im}, X_{in}))$ will thus encode a significant amount of semantic information from $X_{in}$. Since we do not introduce skip connections between the encoder and decoder, we only reconstruct $X_{im}$ from the latent representation $\mathcal{F}_{enc}(\Phi_{modal}(X_{im}, X_{in}))$, which is a challenging task for natural images. However, due to the high structural similarity between different modalities in mpMRI data, with strong contrasts only in certain regions, the cross-modal reconstruction can encourage the model to learn cross-modal representations and explore the correlations between different modalities. Formally, the cross-modal reconstruction loss can be expressed as:

$$\mathcal{L}_{CMR} = ||\mathcal{F}_{dec}(\mathcal{F}_{enc}(\Phi_{modal}(X_{im}, X_{in})) - X_{im}||_2. \quad (2)$$

### 3.2. Modality-wise Data Distillation

The primary objective of the foundation model is to extract highly generalizable latent representations. However, the proxy tasks currently used in pre-training models are often unrelated to the downstream application tasks. We attempt to introduce certain bridging components during the pre-training stage that can guide the pre-training process in acquiring the necessary specific representations. Simultaneously, we hope that these bridging components can facilitate the feature expression of the pre-trained model when

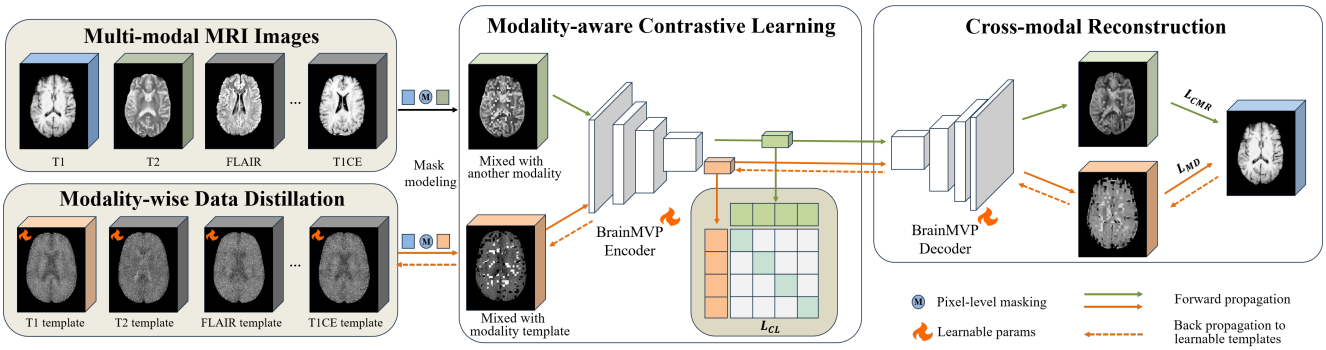



Figure 2. Overview of the proposed BrainMVP, comprised of (a) cross-modal reconstruction module that aims at learning a mapping from images masked with another modality to the original; (b) modality-wise data distillation module that learns condensed modality templates via gradient backpropagation; and (c) modality-aware contrastive learning module for introducing study/case-level modality invariance to the learned features.

applied to downstream tasks. As shown in Fig. 2, modality-wise data distillation is in conjunction with the cross-modal reconstruction process. Specifically, in the cross-modal reconstruction part, we use data either from another modality image $X_{in}$ to fill in the masked region in $X_{im}$ or from the corresponding learnable modality template.

Specifically, the learnable modality templates $T = \{T_m\}_{m=1}^{S}$ sized $S \times H \times W \times D$ are initialized with zero, where $S$ represents the number of modalities in the pre-training datasets. Similar to cross-modal reconstruction, the image needed for filling in $X_{im}$ is $T_m$ ($m$ represents the corresponding modality) instead of another modality, and the remaining steps are the same. An example of learned modality templates is shown in Fig. 4, which demonstrates a compact representation of the structural information for each modality along the pre-training trajectories. Given the masking strategy for modality-wise data distillation denoted as $\Phi_{distill}$, the corresponding loss can be expressed as:

$$\mathcal{L}_{MD} = ||\mathcal{F}_{dec}(\mathcal{F}_{enc}(\Phi_{distill}(X_{im}, T_m)) - X_{im}||_2 . \quad (3)$$

Cross-modal reconstruction and modality-wise data distillation are performed simultaneously. The model needs to learn not only the structural information of a specific modality to form the distilled modality templates but also the transformation relationship between modalities. The representations learned by our pre-trained model are considered modality-agnostic and contain fused representations of different modalities.

### 3.3. Modality-aware Contrastive Learning

As described in section 3.1 and 3.2, $\Phi_{modal}(X_{im}, X_{in})$ and $\Phi_{distill}(X_{im}, T_m)$ still contain $(1 - p^*)$ proportion of information about $X_{im}$, we aim to keep feature-level consistency. To this end, we use contrastive loss to close the high-dimension feature discrepancy. Given their partial semantic consistency, $\Phi_{modal}(X_{im}, X_{in})$ and $\Phi_{distill}(X_{im}, T_m)$ form positive pairs. This can be formalized as:

$$\mathcal{L}_{f_{im} \to g_{im}} = -\log \frac{\exp(f_{im} \cdot g_{im}^T / \tau)}{\sum_{j=1}^{|\mathcal{B}|} \exp(f_{im} \cdot g_{jm}^T / \tau)}, \quad (4)$$

where $f_{im}$ represents the embedding from the current modality image masked with another in the same study, while $g_{im}$ represents the embedding from the current modality image masked with the corresponding distilled template. $|\mathcal{B}|$ denotes the number of positive pairs in a batch. The loss $\mathcal{L}_{g_{im} \to f_{im}}$ is calculated by swapping $f_{im}$ and $g_{im}$. The total loss is the sum of both terms:

$$\mathcal{L}_{CL} = \frac{1}{2} \left( \mathcal{L}_{f_{im} \to g_{im}} + \mathcal{L}_{g_{im} \to f_{im}} \right). \quad (5)$$

**Overall loss.** In summary, the total loss for the proposed multi-modal self-supervised learning scheme is a combination of $\mathcal{L}_{CMR}$, $\mathcal{L}_{MD}$, and $\mathcal{L}_{CL}$:

$$\mathcal{L}_{SSL} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \frac{1}{M_i} \sum_{m=1}^{M_i} (\mathcal{L}_{CMR} + \lambda_{MD} \cdot \mathcal{L}_{MD} + \lambda_{CL} \cdot \mathcal{L}_{CL}), \quad (6)$$

where $\lambda_{MD}$ as well as $\lambda_{CL}$ are coefficients for balancing the corresponding loss term contributions, and $M_i$ denotes the number of modalities of study/case $i$.

### 3.4. Modality templates for downstream application

The distilled modality templates carry shared structural representations of specific modalities from pre-training datasets. We aim to apply these templates to downstream tasks for enhancing generalization performance. In essence, we randomly replace the multi-modal MRI scans of downstream tasks with corresponding distilled templates, aiming to improve the model's modality-invariant representation learning. For detailed implementation, please refer to the supplementary materials.

# 4. Experiments

## 4.1. Datasets

**Pre-training datasets**: We curate a large-scale pre-training dataset collected from five publicly available mpMRI datasets with various sites and acquisition protocols, spanning eight modalities with a total of 3,755 cases/patients containing 16,022 3D scans as demonstrated in Table 1. Among them, the BraTS2021 [2], BraTS2023-SSA [1], and BraTS2023-MEN [29] datasets encompass four common multi-modal brain MRI scans, including T1, T1CE, T2, and FLAIR. The UCSF-PDGM [6] dataset includes 501 cases with additional DWI and ADC modalities. The IXI dataset is utilized as a supplement to the pre-training dataset with cases of normal brains, also incorporating richer modalities such as MRA and PD. Notably, we do not use the segmentation annotations provided in these datasets, and the data do not overlap with the test sets of downstream tasks. More details are in the supplementary materials.

**Downstream datasets**: Ten datasets with various MR imaging sequences are employed for evaluation. These include segmentation tasks: (1) pediatric tumor segmentation BraTS2023-PED [26]; (2) brain metastases segmentation BraTS2023-MET [37]; (3) ischemic stroke lesion segmentation ISLES22 [22]; (4) brain structure segmentation MRBrainS13 [36]; (5) gliomas segmentation UCSF-PDGM [6]; (6) vestibular schwannoma segmentation VSseg [41]; and classification tasks: (1) high-grade and low-grade glioma classification BraTS2018 [3]; (2) mild cognitive impairment classification ADNI [23]; (3) attention deficit hyperactivity disorder classification ADHD-200 [11]; (4) autism spectrum disorder classification ABIDE-I [14]. More details about these downstream datasets are in the supplementary materials.

| Dataset | Task type | Modality type | cases |
|---|---|---|---|
| ***Pre-training*** | | | 3755 |
| BraTS2021 [2] | - | T1,T1CE,T2,FLAIR | 1470 |
| BraTS2023-SSA [1] | - | T1,T1CE,T2,FLAIR | 75 |
| BraTS2023-MEN [29] | - | T1,T1CE,T2,FLAIR | 1141 |
| UCSF-PDGM [6] | - | T1,T1CE,T2,FLAIR,DWI,ADC | 501 |
| IXI | - | T1,T2,MRA,PD | 568 |
| ***Downstream*** | | | |
| BraTS2023-PED [26] | seg. (pediatric tumor) | T1,T1CE,T2,FLAIR | 99 |
| BraTS2023-MET [37] | seg. (brain metastases) | T1,T1CE,T2,FLAIR | 238 |
| ISLES22 [22] | seg. (ischemic stroke lesion) | FLAIR,DWI,ADC | 238 |
| MRBrainS13 [36] | seg. (CF,GM,WM) | T1,T1CE,FLAIR | 20 |
| UPENN-GBM [4] | seg. (glioblastoma) | T1,T1CE,T2,FLAIR | 127 |
| VSseg [41] | seg. (vestibular schwannoma) | T1 | 242 |
| BraTS2018 [3] | cls. (HGG and LGG) | T1,T1CE,T2,FLAIR | 285 |
| ADNI [23] | cls. (MCI and NC) | T1 | 1348 |
| ADHD-200 [11] | cls. (ADHD and NC) | T1 | 767 |
| ABIDE-I [14] | cls. (ASD and NC) | T1 | 819 |

Table 1. Details of datasets used in our work. seg.: segmentation; cls.: classification; CF: Cerebrospinal Fluid; GM: Gray Matter; WM: White Matter; HGG: Higher Grade Glioma; LGG: Lower Grade Glioma; MCI: Mild Cognitive Impairment; NC: Normal Control; ADHD: Attention Deficit Hyperactivity Disorder; ASD: Autism Spectrum Disorder.

https://brain-development.org/ixi-dataset/

## 4.2. Implementation details.

We adopt UniFormer [31] as the backbone of BrainMVP due to its natural multi-modal fusion capabilities. In addition, we have conducted experiments based on the UNET3D [40] network as well. All the experiments are implemented with PyTorch and are run on 8 NVIDIA GeForce RTX 4090 GPUs. Referring to [33], we set $r = 8$ and $p^* = 0.875$. $\lambda_{MD}$ and $\lambda_{CL}$ are both set to 1.0 for equal treatment. During pre-training, we use the AdamW [35] optimizer with a momentum of 0.9 and the weight decay is 1e-5. We train the model for 1,500 epochs with a batch size of 3 and introduce the modality-aware contrastive learning module at epoch 1000 (when the distilled templates have been trained visually well and the corresponding loss has converged, shown in Fig.4). The initial learning rate is set to 3e-4 and we employ a cosine learning rate decay strategy. Detailed hyperparameters for downstream experiments can be found in the supplementary materials.

**Comparison methods**. We compare our BrainMVP against three different types of approaches, i.e., training from scratch, general domain SSL methods, and medical domain SSL methods. There are three mainstream medical image segmentation networks for training from scratch: UNETR [19], UNET3D [40], and Swin-UNETR [18]. UniFormer [31] is a novel 3D medical image segmentation network initially developed in the field of video object detection and extensive experiments have been conducted to verify its effectiveness. The subsequent SSL methods are pre-trained on the above architectures, allowing for a fair comparison of the impact of different network architectures on the final performance. The baseline SSL methods include MAE3D [10, 20], MIM-based SimMIM [55], and contrastive learning related MoCoV3 [8] for general domain, and MG [65], TransVW [16], GVSL [21], Swin-UNETR [46], and VoCo [53] for medical domain. Specifically, two MIM-based methods in medical domain, namely, DAE [48] and M$^3$AE [33], are also taken for comparison. For MRI modality, we **re-implement** the baseline methods on our pre-training dataset for a fair comparison.

**Label efficiency experiments**. To validate if our BrainMVP, pre-trained on large-scale mpMRI datasets, can significantly reduce annotation workload in clinical practice, particularly for handling label-deficient segmentation tasks (which incur higher annotation costs), we conduct label efficiency experiments on five segmentation and one classification datasets. Specifically, we randomly split the training labeled samples into five partitions and gradually increase the training set size by one partition at a time until reaching the full dataset size. The resulted experiments are configured with 20%, 40%, 60%, 80%, and 100% of the total training data. The validation and test sets are kept the same for a fair comparison. For the comparison methods, we select representative approaches for each pre-training

| Method | Modality | Network | BraTS2023-PED [26] | | | | BraTS-MET [37] | | | | ISLES22 [22] | MRBrainS13 [36] | | | | VSseg [41] | UPENN-GBM [4] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ET | TC | WT | AVG | ET | TC | WT | AVG | IS | CF | GM | WM | AVG | VS | ET | TC | WT | AVG |
| ***From Scratch*** | | | | | | | | | | | | | | | | | | | | |
| UNETR [19] | - | - | 46.46 | 76.43 | 78.66 | 67.19 | 54.01 | 54.87 | 59.44 | 56.11 | 74.65 | 67.55 | 78.73 | 83.69 | 76.66 | 70.28 | 83.10 | 80.88 | 81.98 | 81.99 |
| UNET3D [40] | - | - | 47.12 | 81.60 | 83.94 | 70.89 | 56.44 | 58.75 | 62.76 | 59.32 | 80.94 | 70.47 | 73.93 | 82.96 | 75.78 | 69.43 | 85.65 | 88.76 | 86.27 | 86.89 |
| UniFormer [31] | - | - | 46.73 | 83.87 | 86.97 | 72.52 | 67.22 | 72.74 | 70.78 | 70.25 | 84.97 | 77.66 | 74.09 | 75.60 | 75.78 | 80.33 | 87.93 | 91.86 | 88.81 | 89.53 |
| Swin-UNETR [18] | - | - | 49.66 | 81.10 | 84.13 | 71.63 | 63.84 | 67.08 | 68.58 | 66.50 | 75.88 | 70.35 | 81.66 | 84.65 | 78.89 | 76.82 | 87.60 | 91.15 | 87.34 | 88.70 |
| ***With General SSL*** | | | | | | | | | | | | | | | | | | | | |
| MAE3D [10, 20] | Natural | UNETR | 46.55 | 77.08 | 79.32 | 67.65 | 57.45 | 59.19 | 62.06 | 59.57 | 70.43 | 68.30 | 80.57 | 84.69 | 77.86 | 69.57 | 83.66 | 80.42 | 81.86 | 81.98 |
| SimMIM [55] | Natural | UNETR | 45.14 | 76.59 | 78.61 | 66.78 | 54.46 | 55.84 | 58.89 | 56.40 | 69.94 | 68.11 | 80.49 | 84.76 | 77.79 | 69.08 | 83.70 | 81.68 | 82.44 | 82.61 |
| MoCov3 [8] | Natural | UNETR | 45.66 | 77.37 | 79.88 | 67.64 | 55.84 | 56.77 | 61.62 | 58.07 | 70.32 | 67.97 | 79.64 | 84.36 | 77.32 | 69.83 | 83.02 | 80.54 | 81.77 | 81.78 |
| ***With Medical SSL*** | | | | | | | | | | | | | | | | | | | | |
| MG [65] | CXR, CT | UNET3D | 47.99 | **86.69** | 88.41 | 74.36 | 60.11 | 64.05 | 65.43 | 63.19 | 83.53 | 71.40 | 74.71 | 80.41 | 75.51 | 76.33 | 86.64 | 90.58 | 87.03 | 88.08 |
| TransVW [16] | CT | UNET3D | 46.38 | 80.05 | 81.98 | 69.47 | 56.10 | 58.69 | 62.81 | 59.20 | 80.24 | 68.92 | 80.53 | 83.70 | 77.72 | 71.76 | 85.95 | 89.51 | 86.91 | 87.46 |
| GVSL [21] | CT | UNET3D | 49.05 | 84.47 | 86.81 | 73.45 | 62.46 | 66.81 | 67.26 | 65.51 | 80.05 | 69.34 | 75.07 | 82.85 | 75.75 | 72.21 | 87.09 | 91.75 | 87.53 | 88.79 |
| Swin-UNETR* [46] | MRI | Swin-UNETR | 49.07 | 81.74 | 84.13 | 71.65 | 60.60 | 64.56 | 64.53 | 63.23 | 79.55 | 69.67 | **82.09** | **86.13** | 79.30 | 75.55 | 87.24 | 91.46 | 87.28 | 88.66 |
| VoCo [53] | MRI | Swin-UNETR | 48.66 | 82.26 | 84.64 | 71.85 | 57.49 | 59.33 | 63.59 | 60.13 | 77.58 | 71.29 | 76.43 | 81.40 | 76.37 | 76.45 | 86.65 | 90.54 | 87.34 | 88.18 |
| DAE [48] | MRI | Swin-UNETR | 49.30 | 82.12 | 84.78 | 72.07 | 62.27 | 65.99 | 64.85 | 64.37 | 73.92 | 71.37 | 78.50 | 83.20 | 77.69 | 74.51 | 86.90 | 90.83 | 87.32 | 88.35 |
| $M^3$AE [33] | MRI | UNET3D | 46.77 | 85.67 | 86.89 | 73.11 | 66.01 | 70.92 | 70.18 | 69.04 | 83.85 | 71.32 | 69.56 | 79.28 | 73.39 | 75.96 | 87.15 | 91.90 | 88.44 | 89.16 |
| $M^3$AE [33] | MRI | UniFormer | 50.77 | 84.95 | 86.70 | 74.14 | 68.08 | 72.35 | 70.74 | 70.39 | 86.32 | 78.23 | 77.20 | 76.43 | 77.29 | 79.31 | 87.75 | 92.43 | 88.72 | 89.63 |
| **BrainMVP** | MRI | UNET3D | 47.75 | 85.99 | **88.46** | 74.07 | 67.24 | 71.27 | 68.63 | 69.05 | 83.31 | 68.88 | 74.60 | 82.66 | 75.38 | 76.02 | 87.30 | 91.87 | 88.98 | 89.38 |
| **BrainMVP** | MRI | UniFormer | **55.45** | 86.54 | 88.41 | **76.80** | **70.70** | **75.80** | **74.52** | **73.67** | **86.60** | **81.04** | 78.17 | 81.61 | **80.27** | **83.64** | **88.49** | **92.48** | **89.07** | **90.01** |

Table 2. Experimental results on six downstream **segmentation** datasets. We report the mean Dice score (%) on each dataset and the best results are bolded. The second best results are underlined. CXR: Chest X-Ray; ET: enhancing tumor; TC: tumor core; WT: whole tumor; AVG:average; IS: Ischemic Stroke; CF: Cerebrospinal Fluid; GM: Gray matter; WM: White matter; VS: Vestibular schwannoma.

data modality (natural, CT, and MRI), including MAE3D [10, 20], GVSL [21], MG [65], and VoCo [53]. Notably, we observe that MG [65] exhibits strong generalization performance across many datasets, so we include it for comprehensiveness of the comparison.

**Evaluation metrics**. For segmentation tasks, we use Dice Score and Hausdorff distance at 95th percentile (HD95) as evaluation metrics. For classification tasks, we report accuracy (ACC), area under the curve (AUC), and F1 score for comprehensive assessment with higher metric values indicating better classification performance. Note that the HD95 results and qualitative experimental results are presented in the supplementary materials.

### 4.3. Experiments on downstream tasks

**Superior performance on tumor segmentation datasets.** We first validate our BrainMVP on BraTS2023-PED [26] and UPENN-GBM [4]. As shown in Table 2, medical-specific SSL methods consistently outperform general SSL approaches, as models pre-trained on natural images generalize poorly to medical imaging. Specifically, the best average Dice Score achieved by general SSL methods based on MIM is 67.65%, which is 9.15% lower than BrainMVP's best result of 76.80%. Also, MoCoV3 [8] performs less effectively, achieving 9.16% lower in Dice Score compared to BrainMVP. This disparity arises because typical pre-training methods developed primarily for 2D image tasks often require full images or large patches as input, which is usually impractical for 3D medical images. Our BrainMVP also outperforms medical SSL methods based on mask modeling, such as $M^3$AE [33] (76.80% *vs.* 74.14%) and DAE [48] (76.80% *vs.* 72.07%). We further validate the effectiveness of BrainMVP on UPENN-GBM [4], as shown in Table 2. BrainMVP achieves an average Dice Score of 90.01% and outperforms state-of-the-art methods. **Performance improvement on normal brain structure segmentation dataset.** We utilize the MRBrainS13 [36] dataset for the segmentation of normal brain structures to assess the efficacy of BrainMVP in scenarios with limited normal brain structure cases during pre-training. As detailed in Table 2, our BrainMVP achieves an average Dice Score of 80.27%. In contrast, MG [65], employing multiple proxy tasks, attains 75.51%, and VoCo [53], leveraging position prediction, achieves 76.37%. Based on the UniFormer [31] architecture, BrainMVP surpasses all previous methods and demonstrates a notable 4.49% average Dice Score improvement over training from scratch.

**Strong generalization performance on Unseen datasets.** Given that our pre-training datasets primarily include normal brain structures and those afflicted with glioma, we aim to verify the generalization capabilities of BrainMVP on other types of diseases. To assess this, we evaluate our BrainMVP on three datasets: BraTS-MET [37], ISLES22 [22], and VSseg [41]. For the BraTS-MET [37] dataset focusing on brain metastasis subregion segmentation, as seen in Table 2, our BrainMVP achieves an average Dice Score of 73.67%. Further, BrainMVP notably outperforms existing state-of-the-art methods in medical applications, including MG [65] (63.19%), and Swin-UNETR* [46] (63.23%). In the context of the ISLES22 [22] ischemic stroke segmentation task, which involves abnormalities distinct from tumors targeted in pre-training, BrainMVP achieves substantial improvement compared to MG [65] (86.60% *vs.* 83.53%) and GVSL [21] (86.60% *vs.* 80.05%). For the VSseg [41] dataset focusing on vestibular schwannoma segmentation task, in previous methods,

| Method | Modality | Network | BraTS2018 [3] | | | ADNI [23] | | | ADHD-200 [11] | | | ABIDE-I [14] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ACC | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 |
| ***From Scratch*** | | | | | | | | | | | | | | |
| UNETR [19] | - | - | 0.7895 | 0.7817 | 0.6621 | 0.5672 | 0.6066 | 0.5645 | 0.6688 | 0.6523 | 0.6204 | 0.6121 | 0.5478 | 0.5507 |
| UNET3D [40] | - | - | 0.7368 | 0.7373 | 0.4242 | 0.5756 | 0.4966 | 0.3653 | 0.6494 | 0.6798 | 0.4265 | 0.6061 | 0.5059 | 0.4591 |
| UniFormer [31] | - | - | 0.7762 | 0.7719 | 0.6994 | 0.5546 | 0.6343 | 0.5526 | 0.6039 | 0.6387 | 0.5796 | 0.5879 | 0.4433 | 0.4292 |
| Swin-UNETR [18] | - | - | 0.7018 | 0.7143 | 0.6069 | 0.5672 | 0.5853 | 0.5650 | 0.6494 | 0.6950 | 0.6240 | 0.6121 | 0.5530 | 0.5596 |
| ***With General SSL*** | | | | | | | | | | | | | | |
| MAE3D [10, 20] | Natural | UNETR | 0.7018 | 0.6754 | 0.5645 | 0.5756 | 0.5414 | 0.5651 | 0.6169 | 0.6489 | 0.5906 | 0.6061 | 0.4983 | 0.4591 |
| SimMM [55] | Natural | UNETR | 0.7368 | 0.8349 | 0.7077 | 0.6218 | 0.6026 | 0.5446 | 0.6234 | 0.6567 | 0.5790 | 0.5394 | 0.5819 | 0.5318 |
| MoCov3 [8] | Natural | UNETR | 0.7368 | 0.8135 | 0.7304 | 0.6092 | 0.5769 | 0.5996 | 0.6104 | 0.6265 | 0.6007 | 0.5939 | 0.6284 | 0.5890 |
| ***With Medical SSL*** | | | | | | | | | | | | | | |
| MG [65] | CXR, CT | UNET3D | 0.7368 | 0.9286 | 0.4242 | 0.5756 | 0.5496 | 0.3653 | 0.6169 | 0.6980 | 0.6141 | 0.6121 | 0.6266 | 0.5892 |
| TransVW [16] | CT | UNET3D | 0.7368 | 0.7222 | 0.4242 | 0.4958 | 0.6661 | 0.4450 | 0.6818 | 0.7228 | 0.6271 | 0.6424 | 0.5292 | 0.5003 |
| GVSL [21] | CT | UNET3D | 0.7895 | 0.8516 | 0.7286 | 0.5966 | 0.6661 | 0.5959 | 0.6623 | **0.7309** | 0.6565 | 0.6242 | 0.5244 | 0.4701 |
| Swin-UNETR* [46] | MRI | Swin-UNETR | 0.7368 | 0.5032 | 0.4242 | 0.5462 | 0.5517 | 0.5461 | 0.6299 | 0.6437 | 0.5953 | 0.6303 | 0.4993 | 0.3866 |
| VoCo [53] | MRI | Swin-UNETR | 0.7368 | 0.5135 | 0.4242 | 0.5210 | 0.5740 | 0.5207 | 0.6558 | 0.6971 | 0.6413 | 0.5818 | 0.5626 | 0.5466 |
| DAE [48] | MRI | Swin-UNETR | 0.7719 | 0.8151 | 0.7120 | 0.5294 | 0.5666 | 0.5294 | 0.6688 | 0.7129 | 0.6548 | 0.6061 | 0.5173 | 0.5548 |
| $M^3AE$ [33] | MRI | UNET3D | 0.7370 | 0.6984 | 0.5915 | 0.6008 | 0.6338 | 0.6003 | 0.6364 | 0.7049 | 0.6177 | 0.6061 | 0.5453 | 0.4769 |
| $M^3AE$ [33] | MRI | UniFormer | 0.7895 | 0.8659 | 0.7159 | 0.6092 | 0.5352 | 0.5756 | 0.6169 | 0.6597 | 0.6028 | 0.5636 | 0.4682 | 0.4500 |
| **BrainMVP** | MRI | UNET3D | 0.7895 | 0.7746 | 0.6621 | 0.6555 | 0.6669 | 0.6421 | 0.6818 | 0.7245 | 0.6665 | **0.6970** | 0.5817 | **0.6327** |
| **BrainMVP** | MRI | UniFormer | **0.8596** | **0.9452** | **0.8324** | **0.6765** | **0.6964** | **0.6609** | **0.6883** | 0.7249 | **0.6723** | 0.6182 | **0.6329** | 0.5890 |

Table 3. Experimental results on four downstream **classification** datasets. We report the overall accuracy (ACC), area under the curve (AUC) and F1 score on each dataset. The best results are bolded and the second best results are underlined.

| Task | | | BraTS2023-PED [26] | BraTS2018 [3] | | | ADNI [23] | | |
|---|---|---|---|---|---|---|---|---|---|
| Recon. | Distill. | Contrast. | Dice Score (%) | ACC | AUC | F1 | ACC | AUC | F1 |
| ✗ | ✗ | ✗ | 72.52 | 0.7762 | 0.7719 | 0.6994 | 0.5546 | 0.6343 | 0.5526 |
| ✓ | ✗ | ✗ | 75.16 | 0.7895 | 0.8056 | 0.7286 | 0.6261 | 0.6770 | 0.5552 |
| ✓ | ✓ | ✗ | 75.87 | 0.8421 | 0.9032 | 0.8081 | 0.6261 | 0.6835 | 0.6187 |
| ✓ | ✓ | ✓ | **76.80** | **0.8596** | **0.9452** | **0.8324** | **0.6765** | **0.6964** | **0.6609** |

Table 4. Ablation experimental results on BraTS2023-PED [26], BraTS2018 [3] and ADNI [23] datasets. Recon.: cross-modal reconstruction; Distill.: Modality-wise data distillation; Contrast.: modality-aware contrastive learning. Note that cross-modal contrastive learning relies on the presence of both modules aforementioned.

$M^3AE$ [33] achieves the best performance with 79.31% Dice Score, while our BrainMVP outperforms all previous methods with 83.64% Dice Score.

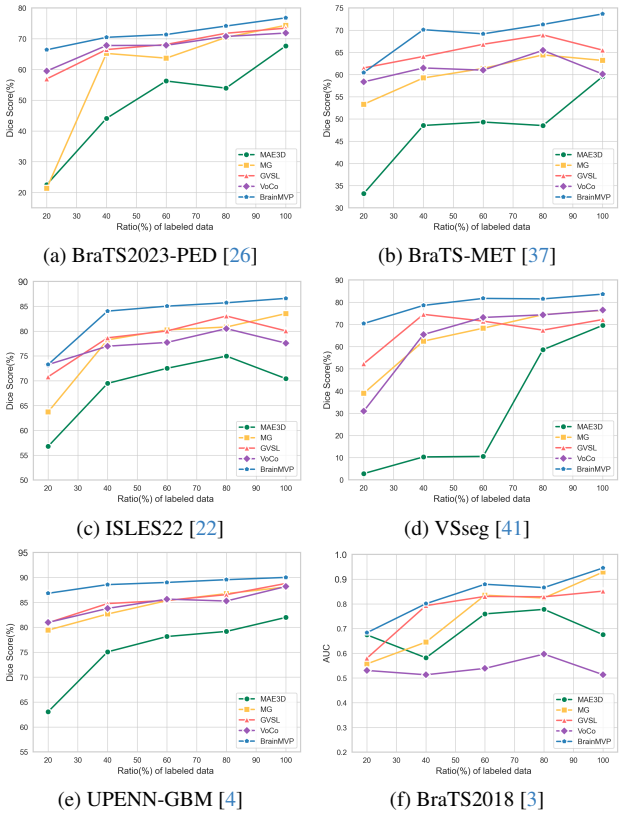
**Classification Results**. We select four distinct classification tasks to assess the generalizability of BrainMVP across diverse domains. As illustrated in Table 3, on the BraTS2018 [3] dataset, our BrainMVP achieves an outstanding ACC of 0.8596, significantly surpassing the state-of-the-art $M^3AE$ [33] (0.7895), VoCo [53] (0.7368), and GVSL [21] (0.7895). BrainMVP also exhibits superior F1 score and AUC compared to prior SSL methods, highlighting its efficacy. Further experiments on ADNI [23], ADHD-200 [11] and ABIDE-I [14] datasets show BrainMVP consistently outperforms state-of-the-art SSL methods. On ADHD-200 [11], BrainMVP achieves an accuracy of 0.6883, surpassing the previous best one of 0.6818. On ABIDE-I [14], BrainMVP improves the accuracy by 5.46%, AUC by 0.45%, and F1 score by 4.35%.

**High Label Efficiency**: Fig. 3 shows that BrainMVP consistently outperforms representative methods when fine-tuned on downstream tasks with varying labeled data ratios. As labeled data increases from 20% to 40%, BrainMVP significantly improves on multiple datasets: BraTS2023-PED [26] (Dice Score 66.41% to 70.46%), BraTS-MET [37] (60.45% to 70.12%), and ISLES22 [22] (73.27% to 84.03%). On BraTS2018 [3], AUC rises from 0.6833 to 0.8008. Notably, with just 40% labeled data, BrainMVP matches or exceeds fully labeled methods. With 20% labeled data, BrainMVP achieves 66.41% Dice Score on BraTS2023-PED [26], 70.39% on VSseg [41], and 86.82% on UPENN-GBM [4], surpassing best-performing methods (59.50%, 52.31%, and 80.97% respectively). This demonstrates BrainMVP's excellent efficiency, reducing annotation needs in clinical practice.

### 4.4. Ablation Study

We perform comprehensive ablation experiments on three key components of BrainMVP: cross-modal reconstruction, modality-wise data distillation, and modality-aware contrastive learning using representative BraTS2023-PED [26], BraTS2018 [3], and ADNI [23] datasets. The results are summarized in Table 4.

**Cross-modal reconstruction**: As shown in Table 4, the inclusion of cross-modal reconstruction in pre-training leads to significant performance improvements. Specifically, on the BraTS2023-PED [26] dataset, the Dice Score increases




(a) BraTS2023-PED [26] (b) BraTS-MET [37]

(c) ISLES22 [22] (d) VSseg [41]

(e) UPENN-GBM [4] (f) BraTS2018 [3]

Figure 3. Label efficiency results of the downstream segmentation and classification tasks. We report the mean Dice Score (%) in segmentation and the area under the curve (AUC) in classification.

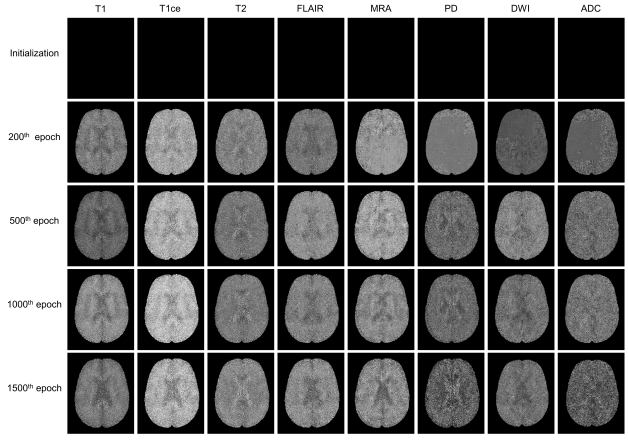



Figure 4. Visualization of distilled modality templates along the pre-training trajectories.

from 72.52% to 75.16%, on the BraTS2018 [3] dataset, the AUC rises from 0.7719 to 0.8056, and on the ADNI [23] dataset, accuracy (ACC) improves from 0.5546 to 0.6261. Notably, for the BraTS2023-PED [26] tumor subregion segmentation task, which requires detailed mpMRI information, the addition of cross-modal reconstruction significantly enhances performance. These results suggest that cross-modal reconstruction effectively captures modality associations, enabling more efficient multi-modal information fusion.

**Modality-wise data distillation**: We next evaluate the effectiveness of the modality-wise data distillation module. As shown in Table 4, the AUC in the BraTS2018 tumor subtype classification task improves significantly, from 0.8056 to 0.9032. Consistent improvement can be seen in BraTS2023-PED [26] and ADNI [23] datasets. This suggests the distilled modality templates learned during pre-training enhance the diversity of downstream data, thereby improving BrainMVP's ability to generalize across tasks.

**Modality-aware contrastive learning**: Finally, we investigate the impact of modality-aware contrastive learning. With its incorporation, BrainMVP's performance consistently improves across multiple datasets. On the BraTS2023-PED [26] dataset, the average Dice Score increases from 75.87% to 76.80%, and on the BraTS2018 [3] dataset for tumor subtype classification, the AUC rises from 0.9032 to 0.9452. For the ADNI [23] dataset, accuracy (ACC) improves from 0.6261 to 0.6765. Modality-aware contrastive learning, supported by cross-modal reconstruction and modality-wise data distillation, contributes to these gains. The combination of these components allows BrainMVP to achieve optimal results, demonstrating the effectiveness of the proposed pre-training framework.

## 5. Conclusion

In this paper, we propose BrainMVP, an efficient multi-modal vision pre-training method for multi-parametric brain MRI analysis. By exploiting structural similarities between MRI modalities, we design cross-modal reconstruction to capture modality correlations. To handle varying numbers of MRI modalities, we use single-channel images, enabling scalability. We also introduce modality-wise data distillation to learn condensed structural representations, and mix input modality images with condensed templates to link pre-training and downstream tasks. Additionally, modality-aware contrastive learning ensures semantic consistency and enhances the model's discriminative ability. Extensive experiments on ten downstream datasets show that BrainMVP outperforms state-of-the-art methods and achieves strong generalizability. Our label efficiency experiment reveals that BrainMVP can match the performance of existing methods using only 40% of labeled data, showcasing its potential for real-world clinical applications.

## Acknowledgments

This work was funded by the National Key R&D Program of China (2022ZD0160700) and Shanghai Artificial Intelligence Laboratory.

## References


[1] Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). *ArXiv*, 2023. 5

[2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 2, 5

[3] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, 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. 2, 5, 7, 8

[4] Spyridon Bakas, Chiharu Sako, Hamed Akbari, M Bilello, A Sotiras, G Shukla, et al. Multi-parametric magnetic resonance imaging (mpmri) scans for de novo glioblastoma (gbm) patients from the university of pennsylvania health system (upenn-gbm). *The Cancer Imaging Archive (TCIA) Public Access*, 2021. 5, 6, 7, 8

[5] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical Image Analysis*, 86:102794, 2023. 1, 2

[6] Evan Calabrese, Javier E Villanueva-Meyer, Jeffrey D Rudie, Andreas M Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T Mongan, and Christopher P Hess. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence*, 4 (6):e220058, 2022. 5

[7] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022. 3

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 5, 6, 7

[9] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 2

[10] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980, 2023. 5, 6, 7

[11] ADHD-200 consortium. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6: 62, 2012. 5, 7

[12] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 3

[13] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. *Advances in Neural Information Processing Systems*, 35:34391–34404, 2022. 3

[14] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014. 5, 7

[15] Yuhang Ding, Xin Yu, and Yi Yang. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3975–3984, 2021. 2

[16] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021. 1, 2, 5, 6, 7

[17] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022. 1, 2

[18] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 5, 6, 7

[19] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 5, 6, 7

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5, 6, 7

[21] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023. 1, 2, 5, 6, 7

[22] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes,

Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 5, 6, 7, 8

[23] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 5, 7, 8

[24] Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15859–15869, 2023. 1

[25] Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z Pan. Delving into effective gradient matching for dataset condensation. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE, 2023. 3

[26] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint arXiv:2305.17033*, 2023. 5, 6, 7, 8

[27] Aishik Konwer, Chao Chen, and Prateek Prasanna. Magnet: Modality-agnostic network for brain tumor segmentation and characterization with missing modalities. In *International Workshop on Machine Learning in Medical Imaging*, pages 361–371. Springer, 2023. 1

[28] Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21415–21425, 2023. 2

[29] Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, et al. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. *arXiv preprint arXiv:2305.07642*, 2023. 5

[30] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Dataset distillation using parameter pruning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 107(6):936–940, 2024. 3

[31] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5, 6, 7

[32] Weibin Liao, Haoyi Xiong, Qingzhong Wang, Yan Mo, Xuhong Li, Yi Liu, Zeyu Chen, Siyu Huang, and Dejing Dou. Muscle: Multi-task self-supervised continual learning to pre-train deep models for x-ray images of multiple body parts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 151–161. Springer, 2022. 1, 2

[33] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. $M^3$ae: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1657–1665, 2023. 1, 2, 3, 5, 6, 7

[34] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *Advances in Neural Information Processing Systems*, 35:13877–13891, 2022. 3

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[36] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015(1):813696, 2015. 5, 6

[37] Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Leon Jekel, Kiril Krantchev, Harrison Moy, Rachit Saluja, Klara Osenberg, Klara Wilms, et al. The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri. *ArXiv*, 2023. 5, 6, 7, 8

[38] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34:5186–5198, 2021. 3

[39] Narinder Singh Punn and Sonali Agarwal. Bt-unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Machine Learning*, 111(12):4585–4600, 2022. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 5, 6, 7

[41] Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, et al. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. *Scientific Data*, 8(1): 286, 2021. 5, 6, 7, 8

[42] Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. $M^2$ftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023. 2

[43] Aron S Talai, Jan Sedlacik, Kai Boelmans, and Nils D Forkert. Utility of multi-modal mri for differentiating of parkinson's disease and progressive supranuclear palsy using machine learning. *Frontiers in Neurology*, 12:648548, 2021. 2

[44] Aiham Taleb, Christoph Lippert, T Klein, and M Nabi. Self-supervised learning for medical images by solving multi-

modal jigsaw puzzles. *Ieee Transactions on Medical Imaging*, 12729:661–673, 2017. 1, 3

[45] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *International conference on information processing in medical imaging*, pages 661–673. Springer, 2021. 3

[46] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022. 1, 2, 5, 6, 7

[47] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 1, 2

[48] Jeya Maria Jose Valanarasu, Yucheng Tang, Dong Yang, Ziyue Xu, Can Zhao, Wenqi Li, Vishal M Patel, Bennett Landman, Daguang Xu, Yufan He, et al. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. *arXiv preprint arXiv:2307.16896*, 2023. 1, 2, 3, 5, 6, 7

[49] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 2

[50] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 2

[51] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022. 3

[52] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 2, 3

[53] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. *arXiv preprint arXiv:2402.17300*, 2024. 1, 3, 5, 6, 7

[54] Yutong Xie, Jianpeng Zhang, Lingqiao Liu, Hu Wang, Yiwen Ye, Johan Verjans, and Yong Xia. Refs: A hybrid pre-training paradigm for 3d medical image segmentation. *Medical Image Analysis*, 91:103023, 2024. 3

[55] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 5, 6, 7

[56] Xiangyi Yan, Junayed Naushad, Chenyu You, Hao Tang, Shanlin Sun, Kun Han, Haoyu Ma, James S Duncan, and Xiaohui Xie. Localized region contrast for enhancing self-supervised learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 468–478. Springer, 2023. 1, 2

[57] Chuyan Zhang, Hao Zheng, and Yun Gu. Dive into the details of self-supervised learning for medical image analysis. *Medical Image Analysis*, 89:102879, 2023. 1

[58] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 3

[59] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 3

[60] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 2, 3

[61] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3499–3509, 2021. 1, 2

[62] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[63] Tongxue Zhou, Su Ruan, and Haigen Hu. A literature survey of mr-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, 104: 102167, 2023. 2

[64] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022. 2

[65] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 1, 2, 5, 6, 7