LARGE-SCALE MODALITY-INVARIANT FOUNDATION MODELS FOR BRAIN MRI ANALYSIS: APPLICATION TO LESION SEGMENTATION

Petros Koutsouvelis[1,†,*], *Matej Gazda*[2,†], *Leroy Volmer*[1], *Sina Amirrajab*[3]
Kamil Barbierik[5], *Branislav Setlak*[2], *Jakub Gazda*[4], *Peter Drotar*[2]

[†]Equal contribution, [*] Corresponding author: petros.koutsouvelis@maastrichtuniversity.nl
[1]Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands
[2]Intelligent Information Systems Lab, Technical University of Kosice, Letna 9, 04201, Kosice, Slovakia
[3] The D-Lab, Department of Precision Medicine, GROW - Research Institute for Oncology and Reproduction, Maastricht University, 6220 MD Maastricht, The Netherlands
[4]2nd Department Of Internal Medicine, Pavol Jozef Safarik University and Louis Pasteur University Hospital, Kosice, Slovakia
[5]SWAI a.s.

ABSTRACT

The field of computer vision is undergoing a paradigm shift toward large-scale foundation model pre-training via self-supervised learning (SSL). Leveraging large volumes of unlabeled brain MRI data, such models can learn anatomical priors that improve few-shot performance in diverse neuroimaging tasks. However, most SSL frameworks are tailored to natural images, and their adaptation to capture multi-modal MRI information remains underexplored. This work proposes a modality-invariant representation learning setup and evaluates its effectiveness in stroke and epilepsy lesion segmentation, following large-scale pre-training. Experimental results suggest that despite successful cross-modality alignment, lesion segmentation primarily benefits from preserving fine-grained modality-specific features. Model checkpoints and code are made publicly available.[1]

***Index Terms*—** SSL, MRI, segmentation, stroke, epilepsy

1. INTRODUCTION

Fueled by recent advances in self-supervised learning (SSL) and public release of large-scale datasets, the field of medical image analysis is shifting toward pre-training general-purpose foundation models [1, 2]. Compared to training from scratch, this paradigm enables (i) learning representations from unlabeled data that may facilitate the identification of subtle abnormalities, (ii) generalizing to various diagnostic tasks and populations with minimal adaptation. These properties are particularly desirable in brain MRI analyses, that remain constrained by the scarcity of clinical-grade annotated cohorts and heterogeneous acquistion protocols.

MRI comprises multiple modalities that offer complementary views of the same underlying biological structure. We hypothesize the existence of subject-specific, modality-invariant representations that can be learned during pre-training by mapping input modalities to a shared latent space. Such a model could capture discriminative structural features while mitigating the impact of missing modalities in retrospective cohorts and reducing future acquisition demands. Existing evidence suggests that unifying cross-modality representations can improve brain lesion segmentation, though performed at smaller pre-training scales [3, 4] or directly during the downstream task [5].

This study evaluates the effectiveness of large-scale foundation model pre-training for downstream lesion segmentation, focusing on stroke and epilepsy. A modality-invariant representation learning framework is introduced and integrated with a masked reconstruction objective, given its demonstrated advantages in segmentation tasks [1]. The contribution of the modality-invariance branch is assessed both independently and in comparison with established SSL approaches. To promote further research, all pretrained foundation models are publicly released.

2. METHODS

2.1. Data Curation and Pre-processing

Pre-training was performed using *FOMO60k* [2], a large-scale brain MRI dataset containing 60,529 scans and multiple modalities across 11,187 subjects, aggregated from 16 sources. To ensure consistency, all images were skull-stripped with HD-BET [6], regardless of their original pre-processing pipeline. The scans were then resampled to isotropic 1 mm^3 resolution, reoriented to RAS coordinate system, and cropped to the brain bounding box. Intensity values were clipped to

[1]https://github.com/BraveDistribution/UMBRA/tree/main

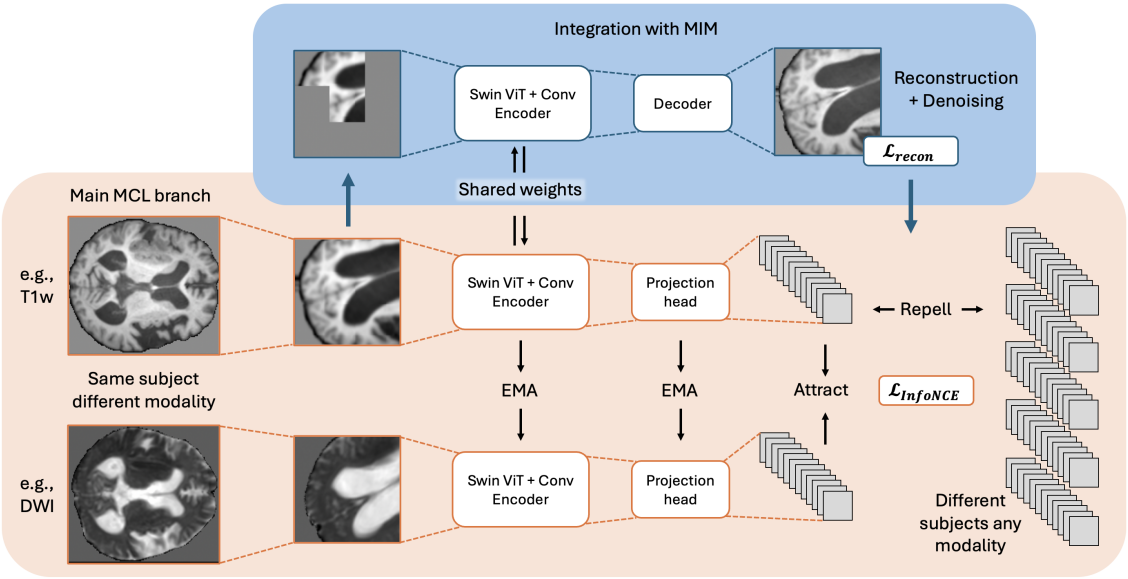

Fig. 1. High-level schematic of the pre-training setup. *Abbreviations*: EMA = exponential moving average.

the 0.5–99.5% percentile range and z-score normalized within the brain mask. Images with any spatial dimension smaller than 12 cm were excluded (n = 8,164), as this empirically determined threshold effectively removed scans with missing brain tissue due to registration artifacts.

2.2. Modality-Invariant Contrastive Learning (MCL)

We employed a *MoCo v2*-based contrastive learning (CL) setup [7], where co-localized 3D brain MRI patches across all scans in a single acquisition session defined the set of *positive pairs*. Patches from different subjects were considered *negative* irrespective of their location and modality, while same-subject different-session comparisons were excluded due to potential longitudinal shifts. Positive pairs underwent separate affine and intensity augmentations to discourage capturing trival relationships. We compared MCL with the default CL implementation, where positive pairs were defined as augmented views of the same modality.

2.3. Integration with Masked Image Modeling (MIM)

Masking was performed at the bottleneck and interpolated back to the patch embedding space to assign masked patches with a learnable token [8]. This setup ensured no leakage of umasked tokens at any stage of the encoder and compatibility with both convolutional and vision transformer (ViT) layers. The target preceded intensity augmentations to encourage noise-invariant representations. To observe emerging dynamics from the two objectives, the MCL branch was replaced with the default CL as an ablation experiment. A high-level schematic of the setup is depicted in Figure 1.

2.4. Implementation Details

A hybrid Swin encoder with residual convolutions [9] was used in all experiments on input volumes of 96 mm^3. The queue size in MCL was set to 16,384 and the *InfoNCE* temperature was fixed at $\tau = 0.2$ [7]. The EMA momentum was linearly increased from 0.996 to 0.999 during the first 20% of training steps. A lightweight feature pyramid network (FPN) decoder mapped features from the MIM branch to the voxel space. The masking ratio varied uniformly between 0.5 and 0.75. Pre-training was performed for 200k steps using the AdamW optimizer (weight decay 0.01) with a base learning rate of 1×10^{-4} and a batch size of 16. Finetuning continued for 5k steps; the FPN decoder was trained from scratch, while the pre-trained encoder remained frozen for the first 30% of steps. No hyperparameter tuning was conducted. Segmentation outputs were obtained via sliding-window inference without further post-processing.

2.5. Evaluation Setup

Downstream lesion segmentation was evaluated on three cohorts: (i) *ATLAS v2* [10], a multi-center dataset of 655 post-stroke cases with T_1-weighted images and expert-annotated lesion masks; (ii) *ISLES 2022* [11], a two-center dataset of 255 ischemic stroke cases with annotations on DWI and accompanying ADC and FLAIR sequences; and (iii) *FCD BONN* [12], a single-center dataset of 85 focal cortical dysplasia (FCD) patients with manual lesion annotations in FLAIR and additional T_1-weighted MRI. All images were rigidly co-registered and pre-processed as *FOMO60k*.

Unique models were trained per imaging modality and tested on held-out sets using the Dice similarity coefficient

(DSC). When feasible, test sets were constructed via center-wise splits; 10 randomly selected centers for *ATLAS v2* (n = 255) and center #2 (n = 52) for *ISLES 2022*. *FCD BONN* distinguishes a subset of 28 challenging cases, which served as the held-out set. Given the absence of additional modalities in *ATLAS v2*, the training set (n = 400) was further partitioned into subsets of 100%, 75%, 50%, and 25% to assess robustness to reduced sample size. Results were compared against training from scratch (i.e., no pre-training) baselines.

3. EXPERIMENTAL RESULTS

The proposed modality-invariant pre-training framework was first evaluated for its ability to align cross-modality embeddings, and subsequently for its impact on downstream lesion segmentation performance.

3.1. Modality Invariance in the Embedding Space

Cross-modality alignment was measured via the mean cosine distance between encoder embeddings from different acquisitions of the same anatomical region. Given the centrally-cropped patch of each subject/session combination in a 2% held-out validation set from *FOMO-60k*, modality-invariant pre-training demonstrated effective cross-modality alignment (MCL: 0.0066 ± 0.0104, MCL + MIM: 0.0068 ± 0.0175) compared to the default counterparts (CL: 0.5995 ± 0.1576, CL + MIM: 0.5028 ± 0.1592). This is further illustrated by the t-SNE plots in Figure 2, where cross-modality embeddings appear considerably more clustered. Cross-subject separability in T_1-weighted inputs remains relatively consistent across frameworks (MCL: 0.6094 ± 0.2500, MCL + MIM: 0.6232 ± 0.2732, CL: 0.6540 ± 0.2023, CL + MIM: 0.6146 ± 0.2198), suggesting no collapse of discriminative features.

3.2. Lesion Segmentation Performance

Segmentation results across all datasets, modalities, and pre-training methods are summarized in Table 1 and Figure 3. Overall, aligning cross-modality representations during pre-training did not steer the models into capturing richer spatial information, highlighted by the comparable performance between MCL and default CL. Instead, performance was mainly driven by the presence of a MIM branch, due to the property of the reconstruction objective to encourage highly-localized embeddings.

Integrating MCL with MIM appeared to deteriorate per-modality performance compared to the default CL with MIM, except in few-shot ablations. This could be attributed to the MCL objective limiting the effectiveness of MIM by (i) yielding higher gradient norms and consequently dominating the optimization process, and (ii) eliminating modality-specific texture features that are necessary to reconstruction. The use of a higher τ value in *InfoNCE*, combined with a token-level modality-agnostic reconstruction objective, such as *iBOT* [13], could circumvent these limitations and surpass the default baseline, while preserving modality invariance.

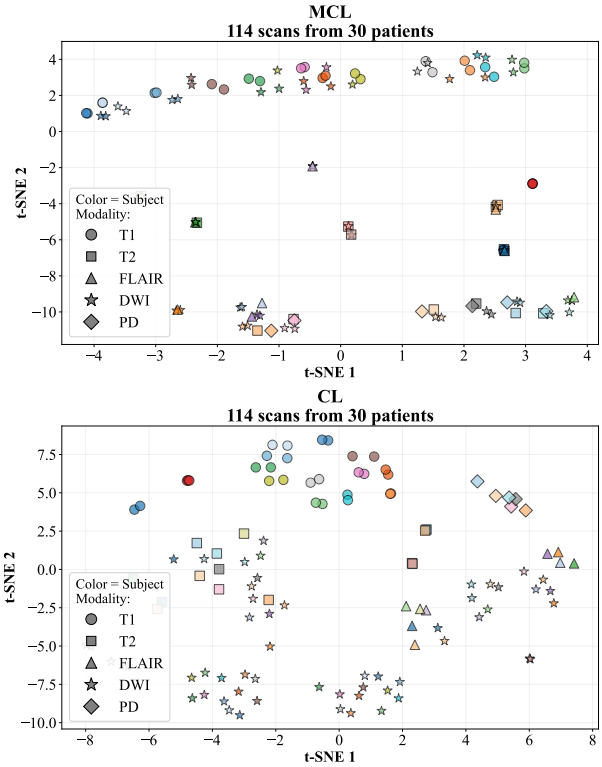

Fig. 2. tSNE on the embeddings from 30 randomly selected cases of the *FOMO60k* validation set using the MCL and CL models.

Table 1. Segmentation DSC with respect to % of training data used in the *ATLAS v2* dataset.

Method	100%	75%	50%	25%
From Scratch	0.4188	0.4339	0.3991	0.3918
CL	0.4678	0.4648	0.4552	0.4377
MCL	0.4693	0.4572	**0.4892**	0.4420
CL + MIM	**0.4936**	0.4872	0.4643	0.4530
MCL + MIM	0.4782	**0.4893**	0.4576	**0.4564**

3.3. Modality Invariance in Downstream Performance

Results in Figure 3 indicate a strong coupling between imaging modality and downstream performance, independent of the pre-training strategy. The highest-scoring modalities were FLAIR for FCD and DWI for stroke lesions, which is concordant with their diagnostic value in clinical practice. This pattern appears unconfounded by the representativeness of the pre-training cohort, as it persists even in models trained from scratch. However, it could be partly attributable to the ground truth annotations being created on the same modalities.

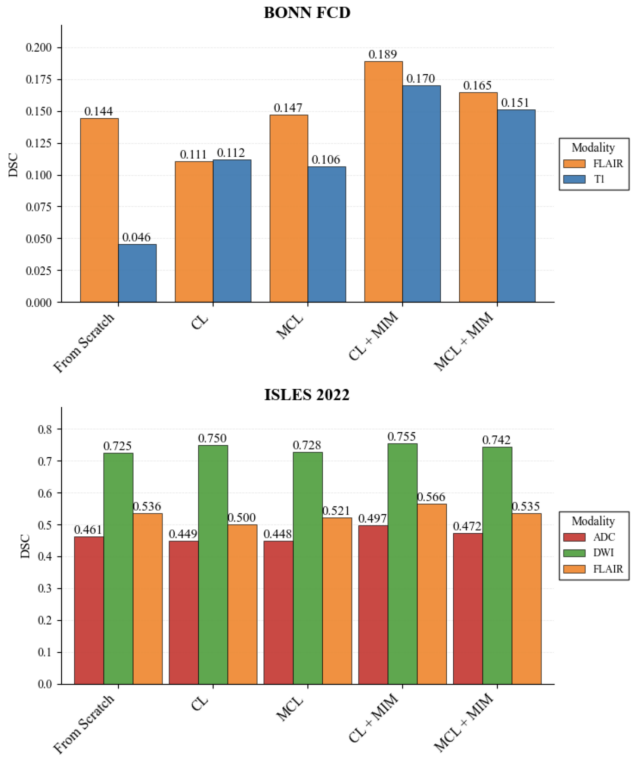

Fig. 3. Segmentation DSC in the *BONN FCD* and *ISLES 2022* datasets for each modality and pre-training method.

Interestingly, pre-trained MCL and MCL + MIM encoders yielded more aligned cross-modality embeddings in the downstream test sets (e.g., in *ISLES 2022* - MCL: 0.1231 ± 0.1126, MCL + MIM: 0.1312 ± 0.1503, CL: 0.6291 ± 0.0816, CL + MIM: 0.6296 ± 0.0804), yet segmentation performance showed similar variation across modalities. This suggests that (i) modality-invariant representations may emphasize global, patient-level anatomical features rather than localized pathological abnormalities, and (ii) each modality captures distinct tissue characteristics that cannot be inferred from other modalities and are hence essential to segmentation. In this regard, the proposed pre-training method could prove more effective for global classification or regression tasks, such as brain age prediction, whereas segmentation studies would benefit from preserving discriminative, modality-specific features and, when feasible, unifying representations through multi-modal fusion.

4. CONCLUSION

This study proposed a modality-invariant self-supervised learning framework and evaluated its effectiveness for lesion segmentation following large-scale pre-training. Experimental results showed that while aligning intra-patient cross-modality representations was feasible, it neither improved per-modality lesion segmentation performance nor reduced dependency on specific modalities. While performance could be improved by optimizing the pre-training setup, the latter may not be feasible, as accurate lesion segmentation relies on fine-grained spatial detail that remains closely tied to modality-specific contrasts and pathology. Future work should investigate the role of modality invariance in global classification or regression tasks, where such representations may prove more beneficial.

5. ACKNOWLEDGMENTS

This work is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00394 and used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-13271.

References

[1] T. Wald *et al.*, "An openmind for 3d medical vision self-supervised learning," 2025.

[2] A. Munk *et al.*, "A large-scale heterogeneous 3d magnetic resonance brain imaging dataset for self-supervised learning," 2025.

[3] L. Chalcroft *et al.*, "Unified 3d mri representations via sequence-invariant contrastive learning," 2025.

[4] Z. Tan *et al.*, "Towards a universal 3d medical multi-modality generalization via learning personalized invariant representation," 2025.

[5] Q. He *et al.*, "Modality-agnostic learning for medical image segmentation using multi-modality self-distillation," in *Proc. ISBI 2024*. IEEE, 2024, pp. 1–5.

[6] F. Isensee *et al.*, "Automated brain extraction of multisequence mri using artificial neural networks," *Human Brain Mapping*, vol. 40, no. 17, p. 4952–4964, Aug. 2019.

[7] X. Chen *et al.*, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[8] Z. Xie *et al.*, "Simmim: A simple framework for masked image modeling," 2022.

[9] Y. He *et al.*, "Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation," in *Proc. MICCAI 2023*. Cham: Springer Nature Switzerland, 2023, pp. 416–426.

[10] S.-L. Liew *et al.*, "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms," *Scientific data*, vol. 9, no. 1, p. 320, 2022.

[11] M. R. Hernandez Petzsche *et al.*, "Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Scientific data*, vol. 9, no. 1, p. 762, 2022.

[12] F. Schuch *et al.*, "An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II," *Scientific Data*, vol. 10, no. 1, p. 475, Jul. 2023.

[13] J. Zhou *et al.*, "iBOT: Image bert pre-training with online tokenizer," 2022.