SELF-SUPERVISED MODALITY-AGNOSTIC PRE-TRAINING OF SWIN TRANSFORMERS

Abhiroop Talasila, Maitreya Maity, U. Deva Priyakumar

Center for Computational Natural Sciences and Bioinformatics International Institute of Information Technology Hyderabad, India

ABSTRACT

Unsupervised pre-training has emerged as a transformative paradigm, displaying remarkable advancements in various domains. However, the susceptibility to domain shift, where pre-training data distribution differs from fine-tuning, poses a significant obstacle. To address this, we augment the Swin Transformer to learn from different medical imaging modalities, enhancing downstream performance. Our model, dubbed SwinFUSE (Swin Multi-Modal Fusion for UnSupervised Enhancement), offers three key advantages: (i) it learns from both Computed Tomography (CT) and Magnetic Resonance Images (MRI) during pre-training, resulting in complementary feature representations; (ii) a domain-invariance module (DIM) that effectively highlights salient input regions, enhancing adaptability; (iii) exhibits remarkable generalizability, surpassing the confines of tasks it was initially pre-trained on. Our experiments on two publicly available 3D segmentation datasets show a modest 1-2% performance trade-off compared to single-modality models, yet significant out-performance of up to 27% on out-of-distribution modality. This substantial improvement underscores our proposed approach's practical relevance and real-world applicability. Code is available at: https://github.com/devalab/SwinFUSE

Index Terms— self-supervision, multi-modal, domain adaptation, 3D image segmentation

1. INTRODUCTION

Supervised deep learning excels at medical image segmentation using lots of labeled data [1, 2, 3]. The shortage of professional radiologists and their limited time and annotation efficiency makes it difficult to get huge medical picture collections with exact annotations. Thus, routine clinical usage of supervised-learning-based segmentation techniques is limited. Recent research has focused on self-supervised learning (SSL), which uses many unlabeled images to learn the general aspects of medical images. Fully supervised model finetuning uses a minimal quantity of labeled data [4]. Effective self-supervised medical image segmentation depends on pretraining quality.



Fig. 1. Visual interpretation of SwinFUSE's attention weights (darker shades indicate higher relevance) for a BraTS21 MRI and the model's segmentation output.

Contrastive learning is a self-supervised pre-training method that minimizes the latent space distance of pairs of similar images (typically produced from the same original image using different data augmentation processes) and maximizes the distance of pairs of dissimilar ones [5]. Contrastive learning methods address domain shift, enhancing model applicability in downstream segmentation networks through consistent data augmentation strategies ensuring similar input distributions. Existing self-supervised learning has two limitations [6].

- Domain Shift: Upstream pre-training uses modified images, affecting downstream segmentation network input distributions. General features from pre-trained models may not apply to segmentation networks.
- Multi-modality: Current techniques often rely on singlemodal data, missing the benefits of multiple modalities. Multi-modal images offer diverse perspectives and augment network segmentation information.

Vision Transformers (ViTs) transformed medical image analysis and computer vision. Transformers thrive in pretext tasks, large-scale training, and layer-based global and local knowledge learning. ViTs simulate long-range global information using self-attention blocks and encode visual representations from patches, unlike Convolutional Neural Networks (CNNs) with small receptive fields. A hierarchical ViT with Shifted Windows (Swin) for local self-attention computing with non-overlapping windows was developed by Liu et al. [7]. Linear architecture has been found to

Corresponding author: deva@iiit.ac.in



Fig. 2. Outline of our proposed pre-training pipeline. Sub-volumes are randomly created from input images and augmented with random inner cutouts and rotations (x_i, x_j) . Each augmentation passes through the patch partition layer to generate embeddings, which are fed to the DIM. The output from the DIM is extracted as kernel densities and forwarded to the Swin Transformer.

be more efficient than ViT's quadratic self-attention layers. Swin UNETR [8] merges feature maps at various sizes using transformer-encoded spatial representations in convolutionbased decoders and achieves state-of-the-art (SOTA) performance in BTCV multi-organ segmentation and Medical Segmentation Decathlon (MSD) challenges [9]. The training paradigm uses proxy activities to learn human anatomical patterns. We extend this intuition to allow complementary feature learning from multiple imaging modalities.

This paper presents SwinFUSE, a modality-invariant selfsupervised pre-training approach for medical image analysis. We utilize Swin UNETR's contrastive learning, masked volume pinpointing, and 3D rotation prediction as proxy tasks for pre-training. Additionally, we introduce a DIM for concurrent feature learning from CT and MRI data. The DIM implicitly identifies relevant input areas and directs them to the Swin Transformer encoder using attention maps (Fig. 1). We train the network on the SynthRad dataset [10] and retain the DIM and encoder for later fine-tuning. Our 3D image segmentation experiments involve the BraTS21 [11] and MSD datasets. We fine-tune the entire network to demonstrate generalization across domains, validating its efficacy for each task, including organ segmentation in the MSD dataset.

2. METHOD

2.1. Datasets

SynthRAD comprises registered brain and pelvis CT images with cone-beam CT and MRI images, serving the purpose of synthetic CT generation for radiotherapy planning [10]. We focus on a subset of 180 patients, utilizing T1weighted gradient-echo MRIs, with some using contrast.

BraTS21 consists of multi-modal MRI scans of glioma, with a total of 1254 patients [11]. The sequences acquired include T1, T2, T1CE, and FLAIR. Segmentation classes include peritumoral edematous/invaded tissue, tumor core, and necrotic tumor core.

MSD contains 2,633 3D images collected from various anatomical regions, modalities, and medical image sources for segmentation purposes [9]. It covers data on body organs or parts like the Brain, Heart, Liver, Lung, Pancreas, Prostate, Hepatic Vessel, Hippocampus, Spleen, and Colon.

2.2. Pre-training

We augment the Swin UNETR architecture using a novel Domain Invariance Module, trained to learn which features to highlight, conditioned on the input type. The training dataset consists of CTs and MRIs; volumes are sampled randomly. During each iteration, 3D patches measuring $x_n \in \mathbb{R}^{96 \times 96 \times 96}$ voxels undergo augmentation with random inner cutout and rotation. These patches are then projected into a C-dimensional space (C = 48) using an embedding layer leading to the DIM as shown in Fig. 2. The two embeddings from respective augmentation are each fed into a 4-layer deep Multi-Head Attention Block (MHA) with 3, 6, 12, 24 heads respectively like Co-Attention [12]. The Q, K, V embedding dimensions increase by an exponent of 2 with the base layer having dimensions $x \in \mathbb{R}^{48}$. The query vector from the first embedding is fed as query input to the

second MHA block and vice versa. Each block in the DIM is initialized with an embedding dimension of 2304. The attention weights are scaled with the original embeddings, and the resulting average embedding is sent as input to the Swin Transformer Encoder. The DIM is constructed as given below, where P denotes the patch partitions being fed into each MHA block (Ψ).

$$DIM: P_1 \cdot \Psi_1(Q_2, K_1, V_1) + P_2 \cdot \Psi_2(Q_1, K_2, V_2) \quad (1)$$

We train the model on the SynthRad dataset using the AdamW optimizer and a warm-up cosine scheduler with 500 iterations on two RTX 3090's. We set the initial learning rate for the pre-training experiments to $4e^{-4}$ and a decay of $1e^{-5}$. We implement our model using PyTorch and MONAI¹.

2.3. Loss Function

We aim to minimize the loss of Swin UNETR's encoder using multiple pre-training objectives, including masked volume inpainting, 3D image rotation, and contrastive coding. Additionally, we maximize an extra loss term that, akin to the approach in [13], employs non-parametric density estimation through kernel density estimation (KDE) and density matching via Jenson-Shannon divergence (JSD). This density-matching loss is a regularizer, ensuring that the feature distribution overlap between the source and target datasets is minimized. The KDE, denoted as $p_{est}(X)$, is formulated as follows:

$$p_{est}(X) = \frac{1}{N} \sum_{n=1}^{N} K\left(\frac{\|X - X_n\|_2}{\sigma}\right)$$
(2)

where $X_1, X_2, X_3, \dots, X_N$ is the number of sampled points from the encoded feature space, the output from the MHA block in our model, and K is a Gaussian kernel. The bandwidth parameter (σ) is estimated to be the mean of the distance between the nearest neighbors in the feature space.

Our loss term for density matching, \mathcal{L}_{JSD} , given density of each MHA block outputs as p_1 and p_2 , is given as follows:

$$JSD_{p_1,p_2} = \frac{1}{2} \{ KL[p_1, M] + KL[p_2, M] \}$$
(3)

where KL is the KL divergence between the two distributions and M is the average of both density estimates. The final loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inpaint}} + \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{rot}} - \mathcal{L}_{\text{JSD}}$$
(4)

2.4. Fine-tuning

In the downstream task, such as 3D image segmentation, we fine-tune the complete Swin UNETR model by removing the projection heads while retaining the DIM. During training, sub-volumes are randomly cropped from the volumetric data.



Fig. 3. Qualitative visualizations of Swin UNETR and our proposed method. Colored regions correspond to necrotic tumor core (red), peritumoral edematous tissue (pink), and enhancing tumor (blue). Dice scores are also given.

Then, stochastic data augmentations, including random rotation and cutout, are applied twice to each sub-volume within a mini-batch, resulting in two different views of each data. All other augmentation parameters align with those used in Swin UNETR.

For SwinFUSE, we utilize pre-trained weights for both the CT and MRI tasks, following the official methods outlined in nnUnet [14] and Swin UNETR [8]. To ensure robustness, we employ a five-fold cross-validation strategy to train models for BraTS21 and MSD experiments. In each fold, we select the best model and ensemble their outputs to generate the final segmentation predictions.

3. RESULTS

Quantitative Average dice scores across five folds for each task in MSD are detailed in Table. 1 for nnUnet, Swin UN-ETR, and our model (with and without pre-trained weights). The proposed method outperforms the current SOTA while segmenting the Brain, Heart, Liver, and Hepatic Vessels but, on average, is 1-2% worse than Swin UNETR. This might be because we reuse the same pre-trained weights to fine-tune

¹https://monai.io/

Organ	Brain	Heart	Hippocampus	Liver	Lung	Pancreas	Prostate	Colon	Hepatic Vessel	Spleen
nnUNET [14]	64.50	94.82	89.76	86.67	72.78	68.31	84.14	59.43	70.14	96.34
Swin UNETR [8]	66.31	94.32	88.39	87.42	76.40	72.91	81.51	60.35	70.75	95.79
SwinFUSE	65.17	94.67	86.76	88.31	74.39	70.95	80.63	60.42	69.78	94.54
SwinFUSE (P)	66.34	94.91	87.91	89.43	75.35	69.25	81.62	59.31	71.61	96.24

Table 1. Average Dice Score across five folds on MSD for two variants of SwinFUSE: with and without pre-training denoted by the presence and absence of (P) respectively.

Table 2. Performance on MRI tasks after pre-training onCT organ regions. Dice scores of Swin UNETR (1) and ourmodel (2) are reported in the format 1/2 in each cell

lested on							
Brain	Heart	Prostate	Hippo-				
Diam	Heart	Trostate	campus				
0.22/ 0.47	0.31/ 0.52	0.19/ 0.32	0.18/ 0.33				
0.21/ 0.29	0.27/ 0.34	0.20/ 0.40	0.16/ 0.26				
0.15/0.32	0.30/ 0.46	0.22/ 0.34	0.23/ 0.30				
0.25/0.32	0.21/ 0.39	0.26/ 0.44	0.25/ 0.48				
0.19/ 0.39	0.24/0.53	0.29/ 0.49	0.27/ 0.38				
0 17/0 35	0 22/0 47	0 20/0 43	0.25/0.41				
0.17/0.35	0.22/0.47	0.29/ 0.43	0.23/0.41				
	Brain 0.22/ 0.47 0.21/ 0.29 0.15/ 0.32 0.25/ 0.32 0.19/ 0.39 0.17/ 0.35	Brain Heart 0.22/0.47 0.31/0.52 0.21/0.29 0.27/0.34 0.15/0.32 0.30/0.46 0.25/0.32 0.21/0.39 0.19/0.39 0.24/0.53 0.17/0.35 0.22/0.47	Brain Heart Prostate 0.22/0.47 0.31/0.52 0.19/0.32 0.21/0.29 0.27/0.34 0.20/0.40 0.15/0.32 0.30/0.46 0.22/0.34 0.25/0.32 0.21/0.39 0.26/0.44 0.19/0.39 0.24/0.53 0.29/0.49 0.17/0.35 0.22/0.47 0.29/0.43				

SwinFUSE for both CT and MRI tasks, differing from Swin UNETR, which uses pre-trained weights for only CT tasks.

Qualitative We visualize images from the MSD (heart, hippocampus, prostate, brain) and BraTS21 datasets in Fig. 3. Segmentation outputs from SwinFUSE are more concise and perform well in the global context. Moderate improvements are seen for smaller organs like the hippocampus (second row). In contrast, for organs like the brain (last row), we notice that the attention mechanism helps locate the unconnected region in the right hemisphere, which Swim UNETR completely fails to detect. In Fig. 1 we generate the learned attention weights from which we can see that the DIM learns similarities between CTs and MRIs and uses those to anchor itself. It further uses differentiating aspects between the both to effectively highlight regions before sending the input to the Swin transformer.

Out-of-distribution When we fine-tune SwinFUSE and Swin UNETR on CT organ regions in MSD like Liver, Lung, and Pancreas, and later test on MRI regions, we notice a significant drop in performance for the latter. For example, when fine-tuned on the Liver and tested on the Brain, Swin UNETR's average dice score is 0.22, whereas SwinFUSE's is 0.47. In another instance, fine-tuning on the Pancreas and testing on the Prostate resulted in Swin UNETR achieving a score of 0.16 compared to 0.26 for ours. Multiple other experiments like these are shown in Table. 2 where we show-

case a minimum improvement of 7% and a maximum of 27% in dice scores. In conclusion, due to the pre-training of our model on a varied collection of human body compositions and its ability to acquire a versatile representation from data obtained from various institutions, we assert that our model is more suitable for clinical applications than existing single-modality models.

4. DISCUSSION AND CONCLUSION

In our research, we have shown noteworthy improvements in the field of medical imaging by using unsupervised pretraining. However, we acknowledge a challenge known as domain shift, where the data used for pre-training differs from the data used for fine-tuning. To address this, we extended the Swin Transformer framework to pre-train SwinFUSE on two distinct medical imaging modalities, CT and MRI. This extension offers three key advantages:

- Complementary Feature Representations: By training on both CT and MRI data, SwinFUSE learns diverse feature representations, making it more adaptable and robust.
- Domain-Invariance Module: Our DIM helps SwinFUSE adapt to domain shifts by emphasizing important regions.
- Remarkable Generalizability: SwinFUSE can perform well on tasks it wasn't initially trained for, making it highly relevant in real-world applications.

Our experiments show that our approach performs slightly worse than single-modality models on in-distribution tasks but significantly outperforms them on out-of-distribution modalities, highlighting its practical applicability. Quantitatively, our method surpassed state-of-the-art models in segmenting the Brain, Heart, Liver, and Hepatic Vessels in the MSD dataset. We emphasize that our diverse pre-training data and versatile representations make SwinFUSE more suitable for clinical use than single-modality models.

We've demonstrated the advantages of pre-training Swin-FUSE using our domain-invariance module and its superior performance on various tasks. We've also highlighted the potential for future research in addressing domain gaps and applying our framework to other medical imaging modalities like PET and X-rays. This work represents a significant advancement in the efficiency and accuracy of medical image analysis, with promising prospects for future research.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data available in open access by RSNA ASNR-MICCAI BraTS 2021, SynthRAD 2023, and Medical Segmentation Decathlon challenges. Ethical approval was not required, as confirmed by the license attached to the open access data.

6. ACKNOWLEDGMENTS

We thank IHub-Data, International Institute of Information Technology, Hyderabad, for their support.

7. REFERENCES

- Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical image analysis*, vol. 41, pp. 40– 54, 2017.
- [2] Zhenghua Xu, Shijie Liu, Di Yuan, Lei Wang, Junyang Chen, Thomas Lukasiewicz, Zhigang Fu, and Rui Zhang, "ω-net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution," *Neurocomputing*, vol. 500, pp. 177–190, 2022.
- [3] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga, "How transferable are self-supervised features in medical image classification tasks?," in *Machine Learning for Health*. PMLR, 2021, pp. 54–74.
- [4] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in *Medical Image Computing and Computer As*sisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. Springer, 2019, pp. 420–428.
- [5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton, "Big selfsupervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22243–22255, 2020.
- [6] Jiaojiao Zhang, Shuo Zhang, Xiaoqian Shen, Thomas Lukasiewicz, and Zhenghua Xu, "Multi-condos: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.
- [7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo,

"Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

- [8] 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*. Springer, 2021, pp. 272–284.
- [9] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al., "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 4128, 2022.
- [10] Adrian Thummerer, Erik van der Bijl, Arthur Galapon Jr, Joost J. C. Verhoeff, Johannes A. Langendijk, Stefan Both, Cornelis (Nico) A. T. van den Berg, and Matteo Maspero, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical Physics*, vol. 50, no. 7, pp. 4664–4674, 2023.
- [11] 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-asnrmiccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.
- [12] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand, "Weakly supervised few-shot object segmentation using coattention with visual and semantic embeddings," arXiv preprint arXiv:2001.09540, 2020.
- [13] Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng, "Unsupervised domain adaptation for medical image segmentation by disentanglement learning and selftraining," *IEEE Transactions on Medical Imaging*, 2022.
- [14] Fabian Isensee, Jens Petersen, Simon AA Kohl, Paul F Jäger, and Klaus H Maier-Hein, "nnu-net: Breaking the spell on successful medical image segmentation," arXiv preprint arXiv:1904.08128, vol. 1, no. 1-8, pp. 2, 2019.