MM-DINOv2: Adapting Foundation Models for Multi-Modal Medical Image Analysis

Daniel Scholz[1,2,3], Ayhan Can Erdur[3,4], Viktoria Ehm[2,5], Anke Meyer-Baese[6,9], Jan C Peeken[4,7,8], Daniel Rueckert[2,3,*], and Benedikt Wiestler[1,2,*]

[1] Chair for AI for Image-Guided Diagnosis and Therapy, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany
[4] Department of Radiation Oncology, TUM University Hospital, Munich, Germany
[5] Chair for Computer Vision and Artificial Intelligence, Technical University of Munich (TUM), Munich, Germany
[6] Department of Scientific Computing, Florida State University, Tallahassee, FL, USA
[7] Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, Munich, Germany
[8] Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Center Munich, Munich, Germany
[9] Institute for Advanced Study, Technical University of Munich (TUM), Munich, Germany
[*]contributed equally as senior authors
daniel.scholz@mri.tum.de

Abstract. Vision foundation models like DINOv2 demonstrate remarkable potential in medical imaging despite their origin in natural image domains. However, their design inherently works best for uni-modal image analysis, limiting their effectiveness for multi-modal imaging tasks that are common in many medical fields, such as neurology and oncology. While supervised models perform well in this setting, they fail to leverage unlabeled datasets and struggle with missing modalities — a frequent challenge in clinical settings. To bridge these gaps, we introduce **MM-DINOv2**, a novel and efficient framework that adapts the pre-trained vision foundation model DINOv2 for multi-modal medical imaging. Our approach incorporates multi-modal patch embeddings, enabling vision foundation models to effectively process multi-modal imaging data. To address missing modalities, we employ full-modality masking, which encourages the model to learn robust cross-modality relationships. Furthermore, we leverage semi-supervised learning to harness large unlabeled datasets, enhancing both the accuracy and reliability of medical predictions. We demonstrate our approach on glioma subtype classification from multi-sequence brain MRI, achieving a Matthews Correlation Coefficient (MCC) of 0.6 on an external test set, surpassing state-of-the-art supervised approaches by +11.1%. Beyond this specific application, our framework provides a scalable and robust blueprint for various multi-modal medical imaging problems effectively leveraging vision foundation

models pre-trained on natural images while addressing real-world clinical challenges such as missing data and limited annotations.[1]

Keywords: DINOv2 · Semi-Supervised Learning · Multi-modal MRI

1 Introduction

Vision foundation models, particularly DINOv2 [20], have demonstrated significant potential in medical image analysis through radiological benchmarks across modalities like MRI, CT, and X-rays [22, 19, 1, 25]. However, existing approaches remain constrained to uni-modal analyses or employ suboptimal multi-modal strategies, such as treating MRI sequences as RGB channels [14]. This limitation prevents their application to clinical tasks requiring joint interpretation of multiple modalities common in fields like oncology or neurology.

Current supervised models for glioma subtype classification achieve competent results when all four standard MRI sequences (T1w, T1ce, T2w, FLAIR) are available [10, 28, 24]. However, these approaches do not utilize large-scale unlabeled datasets like BraTS [2], which contains over 2,000 multi-institutional MRI scans originally curated for segmentation tasks. Further, real-world clinical data often suffers from missing modalities — 15% of patients in routine practice lack at least one essential MRI sequence due to acquisition constraints, protocol variations, or artifacts [21]. Existing methods fail to address this variability, as they either rigidly require fixed modality inputs or process sequences in isolation.

This work addresses these limitations by proposing a novel framework to adapt pre-trained vision foundation models to the specific requirements of multi-modal medical image analysis. Our approach leverages both labeled and unlabeled data while being robust to missing sequences. The contributions of this work are as follows:

1. We propose a novel approach to **adapt pre-trained vision transformers** such as DINOv2 for medical imaging tasks, including a **new multi-modal patch embedding** tailored for **multi-modal imaging data**.
2. We present an adaptive vision transformer architecture that can **handle missing sequences** during training and evaluation. To this end, we extend the existing masking objective with **full modality masking** to encourage the model to learn cross-modality relations.
3. We demonstrate **improved glioma subtype classification** by effectively utilizing large amounts of unlabeled data through semi-supervised learning.

2 Related Work

Deep Learning in Glioma Subtype Classification Glioma subtype classification is crucial for prognosis and treatment planning and has been tackled

[1] The code is publicly available at: `https://github.com/daniel-scholz/mm-dinov2`.

with deep learning-based approaches in many works. Van der Voort et al. [28] combine segmentation and classification tasks to improve glioma subtype classification. Cluceru et al. [7] compare hierarchical classification inspired by genetic markers with standard multi-class classification approaches. The class imbalance in glioma is addressed in [24], where they utilize imbalance-aware supervised loss functions. Ge et al. [11] propose a semi-supervised framework that relies on generative models to impute missing sequences. While these works tackle the glioma subtype classification problem, they do not yet make use of the powerful available foundational models.

Multi-Modal DINOv2 Integrations Foundational models such as DINOv2 have demonstrated strong representation learning in multi-modal settings. [17, 15] focus vision-text integrations but do not address multi-modal imaging. Further, [5] integrates DINOv2 with text data in a medical context for radiology applications. So far, multi-modal imaging with DINOv2 has only been addressed in [14], where they adapt DINOv2 for medical imaging by naïvely stacking modalities as RGB channels, limiting its effectiveness for multi-modal data. Our work more flexibly extends DINOv2 to handle multiple imaging modalities, a common requirement in medical imaging while addressing robustness to missing modalities and leveraging semi-supervised learning.

3 Materials and Methods

Our goal is to enable the use of the pre-trained vision foundation model DINOv2 for multi-modal medical imaging tasks. To this end, we introduce substantial modifications to three key components of DINOv2: the patch embeddings in the vision transformer (ViT) [8] backbone, the masked image modeling, and the image-level objective. An overview of our adaptations is shown in Figure 1.

3.1 Multi-modal Patch Embeddings

ViTs, the backbones of DINOv2, treat an image as a sequence of patches, which are flattened and projected into a sequence of patch embeddings $\mathbf{z}_p$, with $p \in \mathcal{P}$, where $\mathcal{P}$ is the set of patches. The ViTs rely on positional embeddings to encode the order and, hence, the spatial relationships between patches. However, these learned embeddings are designed for uni-modal data, treating all input patches as originating from a single image. Applying this directly to multi-modal images discards the valuable multi-modal information about each input patch. Furthermore, pre-trained vision foundation models lack mechanisms to distinguish input imaging modalities, which is crucial for modality-specific feature extraction, as our experiments show. To address these limitations, we adapt the patch embedding mechanism in the ViT in two ways. First, we make sure the pre-trained positional embeddings are applied separately to the patches of each modality rather than treating all modalities as a single input image. This ensures that the spatial

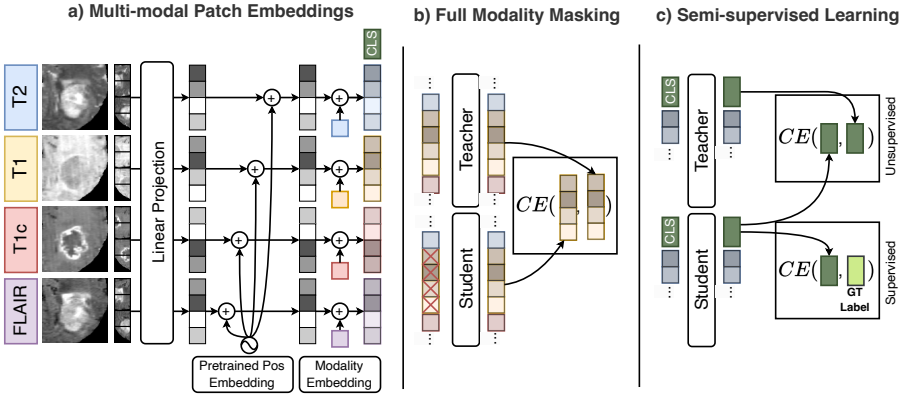

Fig. 1. Schematic representation of our proposed adaptations to DINOv2. To leverage the rich synergies of multi-modal imaging data, we (a) define modality-wise positional and individual modality embeddings (Sec. 3.1), (b) introduce full modality masking to improve robustness against missing sequences (Sec. 3.2), and (c) leverage existing labels in a semi-supervised setup (Sec 3.3).

relationships within each modality are preserved. Formally, we denote the positional embeddings as $\{\mathbf{z}_i\}_{i\in\{1,\dots,|\mathcal{P}|\}}$. These positional embeddings are learned during the DINOv2 pre-training. Second, we introduce modality-specific embeddings. These embeddings are a learnable set of feature vectors $\{\mathbf{z}_m\}_{m\in\mathcal{M}}$ with the set of modalities $\mathcal{M}$. These vectors are initialized as $\mathbf{z}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, yielding the patch embedding $\mathbf{z}_{i,m}$ for the i-th patch in modality m: $\mathbf{z}_{i,m} = \mathbf{z}_p + \mathbf{z}_i + \mathbf{z}_m$. Therefore, the model can efficiently learn modality-specific representations in these sequence embeddings.

3.2 Missing Modality Robustness

In clinical practice, it is common for patients to lack one or more MRI sequences due to acquisition constraints or artifacts. To ensure robustness to such cases, we extend the patch-level objective used in DINOv2. The DINOv2 pre-training employs two ViTs, a student and a teacher, which predict feature representations for each input patch. Input patches to the student are randomly masked while the teacher receives unmasked patches to enforce meaningful representations in the predicted patch representations. A cross-entropy (CE) loss on the masked patches is computed, where the teacher's prediction serves as a target for the student. We leverage this dynamic to encourage robustness to missing sequences by adding full sequence dropout. We mask all patches corresponding to one modality in the student's input sequence while the teacher network receives unmasked patches. The loss remains the same as in DINOv2 [20, 30]. This strategy encourages robust feature learning by forcing the student network

to predict token representations for masked modalities based on cross-modality relationships learned from available sequences.

3.3 Semi-supervised Extension

While DINOv2 is a self-supervised pre-training method, clinical applications often involve datasets with some diagnostic labels available. To leverage these labels during pre-training, we incorporate a semi-supervised mechanism inspired by [9]. Therefore, we extend the image-level objective from DINOv2: The student and teacher receive differently augmented crops of the same input image. Both networks produce image-level probability distributions, *prototype scores*, which are non-linear projections of the CLS token passed through a softmax function. A cross-entropy (CE) loss is calculated between the student's and the teacher's prototype scores, with the latter acting as a pseudo-label: $\mathcal{L}_{image} = -p_t \log p_s$, where p_t and p_s represent teacher and student prototype scores. Since this cross-entropy loss is also often used in supervised learning, it can be easily integrated by replacing the teacher's pseudo-labels with real labels when available. Fini et al. [9] exploit this by formulating a joint loss between the supervised loss and the image-level objective. Here, the prototype score's dimensionality must match the dataset's number of classes. Intuitively, this guides prototype generation so that CLS token clusters align with known class labels.

3.4 Training Setup

We use the ViT-B/14 model pre-trained with DINOv2 as model initialization for all experiments. Since the provided checkpoints do not include weights for the patch- and image-level heads, we randomly initialize these heads and train only them for 10 epochs before unfreezing the entire model. The output size of both heads is set to three to match the number of classes in our dataset. We train the model for 200 epochs on a single NVIDIA A40 or A100 GPU with a batch size of 64 and 42 steps per epoch. The base learning rate is set to 1e−4. Positional embeddings are interpolated to match the size of our input images. Global crops are resized to 98×98 with sizes in the range $(0.5, 1.0)$ of the input image, while local crops are resized to 56×56 with sizes in the range $(0.2, 0.5)$. Crops are always centered around one voxel containing tumor tissue, a step made feasible by automated detection in a standard two-stage workflow; importantly, our approach is not dependent on this cropping and remains applicable without it. For the semi-supervised CE loss, we use a loss weight of 2.0, label smoothing of 0.1, and a temperature scaling of 0.1 in the softmax.

3.5 Dataset

Our dataset comprises preoperative MR images (T1w, T1ce, T2w, FLAIR) from large public datasets of adult patients with newly diagnosed gliomas namely BraTS2021 [18, 3, 2], LUMIERE [26], UPENN GBM [4], Rembrandt [12] UCSF-PDGM [6], EGD [27], and TCGA [3]. We hold out the TCGA dataset [3] for

external testing and randomly split the remaining datasets in 70/10/20% for training, validation, and internal testing. This setup yields 2661 (1162 labeled) subjects for training and validation, 296 for internal, and 214 for external testing. All images provide all four imaging sequences outlined above, while the labeled images have labels from biomarker testing for *IDH* mutation and 1p/19q status in order to classify samples according to the 2021 WHO classification of brain tumors [29] into (a) *IDH* wildtype glioblastoma (GBM), (b) *IDH* mutant and 1p/19q intact astrocytoma (Astro), and (c) *IDH* mutant and 1p/19q codeleted oligodendroglioma (Oligo). The class prevalence in the dataset is 80/10/10%.

All images are resampled to $1 \times 1 \times 1$mm isotropic resolution and rigidly registered to the SRI24 atlas [23]. For training, we randomly sample axial, sagittal, and coronal slices from the volume with at least 500 tumor pixels, crop them to 96×96. We evaluate on 96×96 axial middle slices of the tumor, resized to 224×224, which corresponds to the default evaluation in the DINOv2 code.

4 Results

We rigorously evaluate how adapting pre-trained DINOv2 improves multi-modal medical image classification and enhances robustness to missing modalities. Furthermore, we perform a detailed ablation study to assess each design choice of our proposed method, demonstrating how these contributions collectively improve performance for the clinical application of glioma subtype classification.

4.1 Adapting Vision Foundation Models

To evaluate the effectiveness of our proposed adaptations, we compare two scenarios: fully supervised training and semi-supervised pre-training, followed by linear evaluation. Both scenarios include the following two architectures: (1) RGB DINOv2, a pre-trained DINOv2 with stacked T1ce, T2w, and FLAIR sequences as RGB channels [14] and (2) MM-DINOv2, our multi-modal adaptation of DINOv2. Table 1 summarizes these results. Given the imbalanced nature of the multi-class classification task, we employ Matthews Correlation Coefficient (MCC) as the primary evaluation metric [16], alongside AUROC for classifier calibration and class-wise F1 scores to assess per-class performance.

Supervised Results In the fully supervised setting, only labeled data are used for training (approximately 50% of the full dataset). RGB DINOv2 is compared fully fine-tuned and "frozen", which only utilizes the pre-trained weights. ResNet34 [13] serves as a strong baseline due to its robustness in low-data regimes. Our adapted multi-modal DINOv2 with full fine-tuning outperforms both the ResNet and the RGB DINOv2 variants, indicating the positive influence of our design choices for multi-modal adaptation, as well as the power of pre-trained foundation model features over the randomly initialized ResNet.

Semi-Supervised Results In the semi-supervised setting, all data, including labeled and unlabeled, are utilized. We add the proposed semi-supervised extension (Section 3.3) to RGB and MM-DINOv2. We find our MM-DINOv2 outperforms the RGB DINOv2 in terms of MCC and AUROC. It also further improves in terms of F1 score for two out of three classes compared to the labeled-data-only model, highlighting the strength of incorporating unlabeled images into the training process. Yet, we assume that the unlabeled data introduces more class imbalance compared to the labeled-data-only setting, causing a drop in classification performance in the underrepresented oligodendroglioma class (Oligo).

Table 1. Comparison of supervised and semi-supervised approaches for glioma subtype classification using Matthews Correlation Coefficient (MCC), AUROC, and class-wise F1 scores. Results are reported for RGB DINOv2 with concatenated modalities as RGB channels and our adapted multi-modal DINOv2 (MM-DINOv2) with continuous pre-training. MM-DINOv2 outperforms the comparison methods in all metrics (**best**, best in section).

Method	MCC		AUROC		F1 Score					
					Astro		GBM		Oligo	
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.
Supervised										
ResNet34 [13]	0.58	0.54	0.92	0.79	0.60	0.66	0.92	0.88	0.43	0.35
RGB DINOv2 (frozen) [14, 20]	0.40	0.27	0.87	0.74	0.46	0.31	0.90	0.81	0.35	0.15
RGB DINOv2 [14]	0.55	0.52	0.92	0.80	0.55	0.64	0.90	0.85	0.50	**0.37**
MM-DINOv2 (**ours**)	0.68	**0.60**	**0.95**	**0.89**	0.68	**0.71**	0.94	**0.89**	0.62	0.33
Semi-supervised										
RGB DINOv2 [14]	0.47	0.37	0.84	0.77	0.44	0.42	0.93	0.83	0.40	**0.37**
MM-DINOv2 (**ours**)	**0.74**	0.57	**0.95**	0.86	**0.76**	**0.71**	**0.96**	**0.89**	**0.67**	0.21

4.2 Missing Sequence Robustness

To evaluate the effectiveness of our full modality masking strategy, we compare models trained with and without this design choice on our test sets where one MRI sequence is randomly masked. Across all metrics, including MCC, AUROC, and class-wise F1 scores, the model trained with full sequence masking consistently outperforms the model trained without it. Solely, the external test set performance on the astrocytoma class suffer slightly in terms of F1 score. These results demonstrate that our masking strategy effectively encourages the model to learn cross-modality relationships, enabling robust performance when one modality is missing during inference, a common scenario in clinical practice.

Table 2. Missing sequence robustness analysis. We compare two models trained with all four sequences and either with and without full sequence dropout by evaluating their performance with one sequence randomly missing.

Full Sequence Masking	MCC		AUROC		F1 Score					
					Astro		GBM		Oligo	
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.
No	0.46	0.41	0.86	0.82	0.47	**0.58**	0.92	0.83	0.36	0.13
Yes	**0.57**	**0.46**	**0.89**	**0.83**	**0.61**	0.56	**0.94**	**0.86**	**0.42**	**0.28**

4.3 Ablation Study

We conduct an ablation study to rigorously evaluate the impact of each design choice in our model, ranging from simply concatenating all tokens from all modalities to our full multi-modal adaptation. The results, shown in Table 3, are evaluated in the semi-supervised setting, as it achieved the best overall performance. We observe poor performance initially with a single global positional embedding for all tokens from all modalities treated as a single input modality, which corresponds to spatially concatenating modalities. This is expected, as spatial concatenation implies false spatial correlations between modalities. All our design choices, including modality-specific embeddings (Section 3.1) and full sequence masking (Section 3.2), consistently improve classification performance.

Table 3. Ablation study over our design choices. Adding our proposed adaptations continuously improves the glioma subtype classification performance.

Semi-supervised	MCC		AUC		F1 Score					
					Astro		GBM		Oligo	
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.
Concat Tokens	0.40	0.29	0.84	0.73	0.41	0.31	0.93	0.82	0.29	**0.22**
+ Per-Image Pos Embedding	0.63	0.36	0.93	0.79	0.68	0.45	0.95	0.83	0.35	0.12
+ MRI Sequence Embedding	**0.74**	0.49	0.94	**0.86**	0.71	0.65	**0.97**	0.86	**0.71**	0.13
+ Full Sequence Masking	**0.74**	**0.57**	**0.95**	**0.86**	**0.76**	**0.71**	0.96	**0.89**	0.67	0.21

5 Conclusion

This work introduces a novel adaptation strategy for vision foundation models like DINOv2 to effectively address multi-modal medical imaging challenges. Our approach enhances robustness to missing modalities and leverages partially available labels to improve performance on clinical tasks such as glioma

subtype classification. Our findings highlight the importance of incorporating semi-supervised learning, which outperformed supervised training by utilizing all available data. Moreover, we demonstrated that methodological adaptations tailored to multi-modal medical image analysis are essential for tasks integrating multiple imaging modalities. Notably, our full sequence masking strategy effectively addressed the challenge of missing MRI sequences, a frequent issue in clinical workflows. Our contributions complement existing approaches like Rad-DINO [22] while emphasizing multi-modal data integration and robustness, critical factors for real-world clinical scenarios. Furthermore, our results align with prior studies on semi-supervised learning in medical imaging [11] while extending these insights to foundation models. By systematically addressing the key challenges of multi-modal integration, missing data, and limited annotations, our work provides a generalizable blueprint for adapting vision foundation models to a wide range of multi-modal medical imaging problems. However, our current approach is limited to 2D slice-based training and evaluation, which may not fully capture the spatial relationships inherent in volumetric medical data. Additionally, our experimental setup used cropping strategies that may not be optimal for all clinical scenarios or imaging protocols. Future directions could explore combining our approach with methods like the Medical Slice Transformer [19] to extend foundation models to 3D volumetric imaging tasks, broadening their applicability to radiology and beyond. By addressing critical challenges in multi-modal medical imaging, we hope our work inspires further innovation in diagnostic tools and contributes to improved patient care.

Acknowledgments. This study was supported by the DFG, grant #504320104.

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

References

1. Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., et al.: Evaluating General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 on Radiology Benchmarks (sep 2024)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification (sep 2021)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., et al.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(1), 170117 (sep 2017)
4. Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., et al.: The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics. Scientific Data **9**(1), 453 (jul 2022)
5. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., et al.: MAIRA-2: Grounded Radiology Report Generation (sep 2024)
6. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., et al.: The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. Radiology: Artificial Intelligence **4**(6), e220058 (nov 2022)

7. Cluceru, J., Interian, Y., Phillips, J.J., Molinaro, A.M., Luks, T.L., et al.: Improving the noninvasive classification of glioma genetic subtype with deep learning and diffusion-weighted imaging. Neuro-Oncology **24**(4), 639–652 (apr 2022)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (oct 2020)
9. Fini, E., Astolfi, P., Alahari, K., Alameda-Pineda, X., Mairal, J., et al.: Semi-Supervised Learning Made Simple With Self-Supervised Clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3187–3197 (2023)
10. Foltyn-Dumitru, M., Schell, M., Rastogi, A., Sahm, F., Kessler, T., et al.: Impact of signal intensity normalization of MRI on the generalizability of radiomic-based prediction of molecular glioma subtypes. European Radiology (sep 2023)
11. Ge, C., Gu, I.Y.H., Jakola, A.S., Yang, J.: Deep semi-supervised learning for brain tumor classification. BMC Medical Imaging **20**(1), 87 (jul 2020)
12. Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.C., Fine, H., et al.: The REMBRANDT study, a large collection of genomic data from brain cancer patients. Scientific Data **5**(1), 180158 (aug 2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (dec 2015)
14. Huang, Y., Zou, J., Meng, L., Yue, X., Zhao, Q., et al.: Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification (feb 2024)
15. Jiang, D., Liu, Y., Liu, S., Zhao, J., Zhang, H., et al.: From CLIP to DINO: Visual Encoders Shout in Multi-modal Large Language Models (mar 2024)
16. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., et al.: Metrics reloaded: Recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (feb 2024)
17. Maniparambil, M., Akshulakov, R., Djilali, Y.A.D., Narayan, S., Singh, A., et al.: From Unimodal to Multimodal: Scaling up Projectors to Align Modalities (sep 2024)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (oct 2015)
19. Müller-Franzes, G., Khader, F., Siepmann, R., Han, T., Kather, J.N., et al.: Medical Slice Transformer: Improved Diagnosis and Explainability on 3D Medical Images with DINOv2 (nov 2024)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (jul 2023)
21. Pemberton, H.G., Wu, J., Kommers, I., Müller, D.M., Hu, Y., Goodkin, O., Vos, S.B., Bisdas, S., Robe, P.A., Ardon, H., et al.: Multi-class glioma segmentation on real-world data with missing mri sequences: comparison of three deep learning algorithms. Scientific Reports **13**(1), 18911 (2023)
22. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., et al.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (jan 2025)
23. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Human Brain Mapping **31**(5), 798–819 (dec 2009)

24. Scholz, D., Erdur, A.C., Buchner, J.A., Peeken, J.C., Rueckert, D., et al.: Imbalance-aware loss functions improve medical image classification. In: Medical Imaging with Deep Learning (feb 2024)
25. Song, X., Xu, X., Yan, P.: General Purpose Image Encoder DINOv2 for Medical Image Registration (Feb 2024). https://doi.org/10.48550/arXiv.2402.15687
26. Suter, Y., Knecht, U., Valenzuela, W., Notter, M., Hewer, E., et al.: The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. Scientific Data **9**(1), 768 (dec 2022)
27. van der Voort, S.R., Incekara, F., Wijnenga, M.M.J., Kapsas, G., Gahrmann, R., et al.: The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. Data in Brief **37**, 107191 (aug 2021)
28. van der Voort, S.R., Incekara, F., Wijnenga, M.M.J., Kapsas, G., Gahrmann, R., et al.: Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. Neuro-Oncology **25**(2), 279–289 (feb 2023)
29. Wen, P.Y., Packer, R.J.: The 2021 WHO Classification of Tumors of the Central Nervous System: Clinical implications. Neuro-Oncology **23**(8), 1215 (aug 2021)
30. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., et al.: Image BERT Pre-training with Online Tokenizer. In: International Conference on Learning Representations (oct 2021)