Reconstruct and Distill: A Hybrid Masked-Autoencoder and Self-Distillation Foundation Model for Clinical Brain MRI

Nguyen Quoc Hung, Minh Nguyen Dang, Le Nhu Quynh, Linh Nguyen

Posts and Telecommunications Institute of Technology, Vietnam
nguyenquochung.workvn@gmail.com

Abstract. Self-supervised pretraining on large unlabelled image collections has become a standard way to build transferable encoders for medical imaging, where labelled data are scarce but raw scans are abundant. A brain-MRI encoder pretrained this way is expected to serve a range of downstream clinical tasks, yet these tasks demand different information from it: segmenting a meningioma needs dense local features, while classifying a scan from a frozen embedding needs one globally coherent representation. No single self-supervised objective provides both. We pretrain a 3D residual-encoder U-Net on 302,741 brain-MRI scans jointly with two objectives, masked-autoencoder reconstruction for the local features and DINO-style self-distillation over the pooled bottleneck for the global representation, over a shared set of weights. We then evaluate a single pretrained checkpoint across all seven tasks of the FOMO26 challenge method track, under few-shot and out-of-domain conditions. After only five pretraining epochs and default finetuning it reaches Dice 0.146 on meningioma segmentation, the highest-weighted task type, AUROC 0.602 on polymicrogyria and 0.704 on infarct classification. The reported values are few-shot, out-of-domain scores rather than clinical-accuracy estimates, and absolute Dice is low across all entries. The frozen-embedding and brain-age tasks are weaker, consistent with a self-distillation embedding that continues to improve well beyond five epochs.

Keywords: Foundation models · Brain MRI · Self-supervised learning · Masked autoencoder · Self-distillation.

1 Introduction

Deep learning for brain-MRI analysis has largely followed a supervised paradigm, training a separate model from scratch for each clinical task on manually annotated scans [11]. This paradigm is limited by the cost of annotation: routine imaging produces far more scans than any site can label, and expert annotations are scarce and expensive for most tasks. Self-supervised pretraining addresses this limitation. A single encoder is pretrained once on a large unlabelled corpus and then adapted to each downstream task with the few labelled cases a site

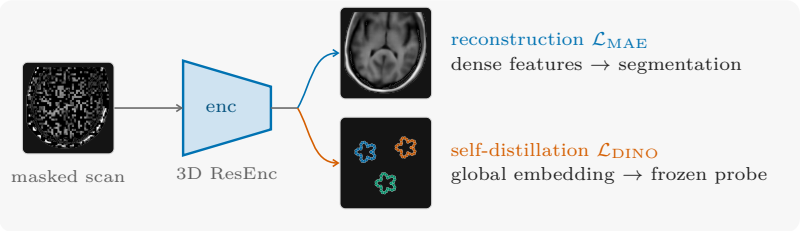

Fig. 1. The two pretraining objectives. A single 3D residual-encoder U-Net encodes a masked scan and is trained by two objectives: masked-autoencoder reconstruction (blue, $\mathcal{L}_{\mathrm{MAE}}$) for the local features used in segmentation, and DINO-style self-distillation over the pooled bottleneck (red, $\mathcal{L}_{\mathrm{DINO}}$) for the global embedding used by the frozen-probe tasks.

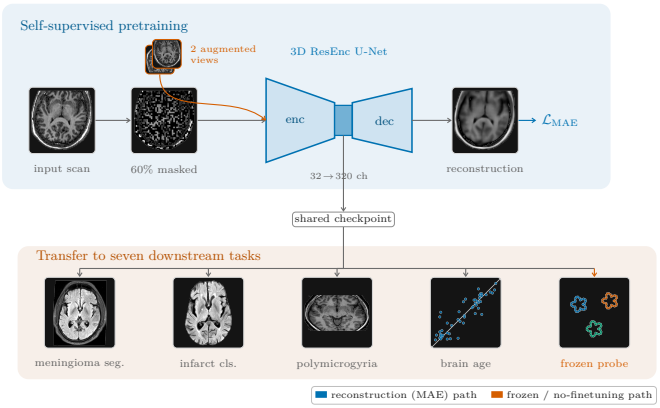

Fig. 2. Overview of the pretraining and transfer pipeline. A single 3D residual-encoder U-Net is pretrained once on 302,741 unlabelled brain-MRI scans: a masked copy of each volume is encoded and reconstructed under $\mathcal{L}_{\mathrm{MAE}}$ (top), while two augmented views are matched under $\mathcal{L}_{\mathrm{DINO}}$. The resulting checkpoint is reused across all seven FOMO26 tasks, finetuned for the image-level and segmentation tasks and frozen for the linear-probe tasks. It reaches Dice 0.146 on meningioma segmentation, the highest-weighted task type, and AUROC 0.602 on polymicrogyria classification. All brain panels are real: the input, mask, and reconstruction are outputs of the pretrained model, and each task thumbnail is a scan from that task.

can assemble [2], which is the foundation-model approach now standard across medical imaging.

A single reused encoder must then serve downstream tasks that demand different information from it. The FOMO26 challenge makes this concrete: on its method track, which restricts pretraining to a shared public corpus, one pretrained checkpoint is scored on seven downstream tasks, from voxel-level meningioma and trigeminal-nerve segmentation, through infarct and malformation classification and brain-age regression, to two tasks that classify directly from the frozen embedding without finetuning. The scoring adds a further con-

straint, because segmentation carries half of the unified score and the frozen-embedding tasks a fifth, so a checkpoint strong at reconstruction but weak at embedding is penalised. The design question is therefore not which single task to optimise, but how to make one encoder carry both dense and global information.

The previous edition of the challenge suggests that no single objective covers this range. Masked-autoencoder reconstruction favoured segmentation, hybrid reconstruction-plus-contrastive objectives favoured classification, and the balance between local and global supervision was left as an open design parameter [11]. Natural images show the same split: masked image modelling yields local features strong for dense prediction and finetuning, whereas contrastive and self-distillation objectives yield globally coherent features strong for linear probing and classification [13]. Combining masked reconstruction with self-distillation in a single encoder is now the standard approach for general-purpose frozen features in natural images [18,12], but no submission to the previous edition applied it in a 3D convolutional encoder. We adopt this pairing for brain MRI.

In this paper we present a 3D brain-MRI foundation model that pretrains one residual-encoder U-Net with two self-supervised objectives, sketched in Figure 1. Figure 2 summarises the full pretraining-and-transfer pipeline and Figure 3 details the architecture. The two objectives target complementary representations. A masked-autoencoder branch reconstructs the volume from a heavily masked copy, which trains the encoder to represent the location of anatomy and abnormality and supplies the dense features segmentation needs. A self-distillation branch, following DINO, matches two augmented views of the same scan and learns a view-invariant summary of the whole volume, which is the global representation the frozen classification tasks use. Both objectives train one encoder jointly, so the model is pretrained once and shared across tasks rather than tuned per task family.

This work makes three contributions. The first is the model itself: a 3D convolutional brain-MRI encoder trained with masked reconstruction and self-distillation together in one backbone. The second is a set of design decisions that adapt the natural-image recipe to clinical MRI, including a single-channel contrast-agnostic input, a lighter 60% mask, and a warmup that ramps the self-distillation loss in only after reconstruction has built a usable feature space. The third is the evaluation: we run one checkpoint on all seven FOMO26 tasks under the challenge's few-shot, out-of-domain protocol, adapting on tens of cases from India or the USA and testing on scans from Denmark. It reaches Dice 0.146 on meningioma segmentation, AUROC 0.602 on polymicrogyria and 0.704 on infarct classification, all from five pretraining epochs and default finetuning, which is consistent with the earlier finding that these models saturate early [11]. The frozen-embedding and brain-age tasks are weaker at this budget. The warmup also leaves a pure masked-autoencoder checkpoint on the schedule, which Section 4.3 uses to measure what self-distillation adds to the pooled embedding.

2 Related Work

Self-supervised learning for 3D medical images. Self-supervised pretraining on unlabelled volumes is now the standard way to build transferable 3D medical encoders. Models Genesis learns anatomy by restoring corrupted patches of chest CT, with features that transfer across organs, diseases, and modalities [19], and masked image modelling on a residual-encoder U-Net is a strong recipe for 3D segmentation [16]. For brain MRI in particular, AMAES pretrains a 3D-native encoder with augmented masked autoencoding and improves downstream segmentation [10]. These methods make reconstruction a reliable local objective, but none adds a global objective, and all are judged mainly on segmentation rather than frozen-embedding transfer.

Local versus global self-supervised objectives. Masked reconstruction and self-distillation learn complementary representations. Masked image modelling captures local, high-frequency structure that helps dense prediction and full finetuning, whereas self-distillation and contrastive objectives capture global, low-frequency structure that helps linear probing and classification [13]. DINO shows that self-distillation with a momentum teacher yields features whose k-nearest-neighbour and linear-probe accuracy is high without any finetuning [1], and alignment-uniformity analysis explains why such an objective spreads classes into a linearly separable arrangement on the embedding sphere [17]. The same objective can also collapse its embedding into a low-dimensional subspace that caps frozen-probe accuracy [8]. Together these results motivate pairing a local and a global objective when one encoder must serve both dense and frozen-embedding tasks.

Hybrid objectives and their absence in 3D brain MRI. Combining masked modelling with self-distillation is the current recipe for general-purpose frozen features in natural images. iBOT unifies masked-patch distillation with image-level self-distillation under a single momentum teacher [18], and DINOv2 scales a combination of the DINO and iBOT losses into a foundation model whose frozen features serve both image-level and pixel-level tasks without finetuning [12]. In the previous brain-MRI challenge, hybrid objectives were built almost entirely from masked reconstruction plus contrastive learning, and no team combined masked reconstruction with self-distillation in a 3D convolutional encoder [11]. We apply this combination to a 3D residual-encoder U-Net for clinical brain MRI.

3 Method

3.1 Overview

We pretrain a single 3D residual-encoder U-Net once on unlabelled brain MRI, reuse it for every downstream task, and train it with a masked-autoencoder re-

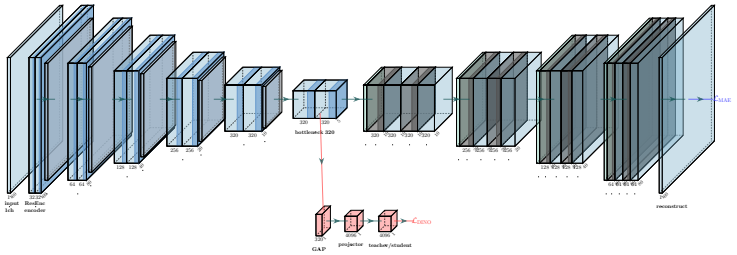

Fig. 3. The hybrid pretraining architecture. A single 3D residual-encoder U-Net (blue) forms the reconstruction branch: the encoder contracts a single-channel volume through six stages (32 to 320 channels), and the decoder expands it back to reconstruct the masked input under $\mathcal{L}_{\mathrm{MAE}}$. The self-distillation branch (red) pools the 320-channel bottleneck and projects it to 4096 prototypes, where a student matches an exponential-moving-average teacher under $\mathcal{L}_{\mathrm{DINO}}$. For downstream use the pretrained decoder is finetuned for segmentation, a pooling head is finetuned for the image-level tasks, and the pooled bottleneck is used directly for the two no-finetuning tasks.

construction loss and a DINO-style self-distillation loss at the same time, combined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{MAE}} + \lambda(t)\,\mathcal{L}_{\mathrm{DINO}}, \tag{1}$$

where $\lambda(t)$ ramps the self-distillation term in after a reconstruction-only warmup.

3.2 Backbone and single-channel input

We use a 3D residual-encoder U-Net [14,5] from the nnU-Net family as the backbone [6,7]. Its encoder has six stages with 32, 64, 128, 256, 320, and 320 channels and residual blocks of depth 1, 3, 4, 6, 6, and 6, with instance normalisation throughout. At 90.3M parameters in the encoder and 97.2M in the encoder-decoder, the model stays in the size range that reached the top of the previous challenge [11] while meeting the challenge's runtime limit of 120 s per case on a single H100. We keep an encoder-decoder U-Net rather than a plain encoder because the decoder is the reconstruction head during pretraining and the segmentation head during finetuning, so the structure that half of the FOMO26 score depends on is trained from the first step.

The encoder reads a single-channel volume and never assumes which MRI sequence it sees. This choice is forced by the data, as Section 4.2 details: a fixed multi-channel layout would be undefined on most scans [2], whereas every scan is a valid single-channel input on its own. At finetuning the same weights handle any contrast, since we repeat the pretrained stem across whatever channels a task provides. We use a convolutional backbone rather than a transformer [3,15], because these encoders dominated the previous challenge and a convolutional hybrid inherits what already works there [11].

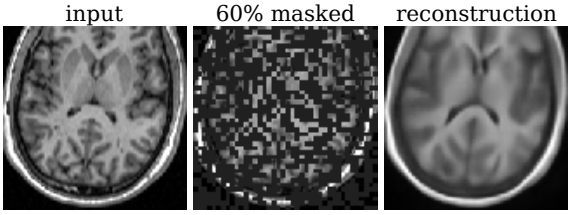

Fig. 4. Masked reconstruction from the pretrained encoder on a held-out T1w scan. Left to right: an input axial slice, the same slice with 60% of the 4^3 tokens masked, and the reconstruction. The reconstruction is smooth but preserves the ventricles, midline, and gray–white boundary. This qualitative example uses the same epoch-5 checkpoint as the rest of the paper and one random mask draw.

3.3 Masked-autoencoder branch

The reconstruction branch builds a spatial prior over the location of anatomy in the volume, which is what the segmentation tasks depend on. A clinical scan is mostly ordinary anatomy with a small abnormal region, so an objective that forces the network to predict missing tissue from its surroundings is a direct fit. We follow the masked-autoencoder recipe [4] and its 3D medical adaptations [10,16], with two changes suited to clinical MRI. We mask 60% of the volume rather than the 75% used on natural images, because a 3D brain scan is more redundant along the slice axis and a lighter mask still leaves the task non-trivial. The masked 4^3-voxel tokens are zeroed, and the whole 160^3 patch is passed through the encoder-decoder. We also score the mean-squared error over the entire volume, not the masked tokens alone, which keeps the intact context in the loss and steadies training on heterogeneous scans. Figure 4 shows the pretrained encoder recovering the ventricles, the midline, and gross tissue structure from a 60% mask.

3.4 Self-distillation branch and objective schedule

A reconstruction-only encoder organises its features for local prediction, and these features need not be linearly separable, which is what the two frozen-embedding tasks require, since they classify from the embedding with no finetuning to reshape it. We add a DINO-style self-distillation objective on the globally pooled encoder bottleneck to supply this property [1]. The objective matches the embeddings of two augmented views of the same scan, so the encoder learns features that are invariant to those augmentations. We build the two views with random flips and light per-view intensity scaling and shifting, bounded to ± 0.1 so that a contrast change never rewrites the anatomy the embedding should preserve. Each view passes through the encoder, its 320-channel bottleneck is pooled by global averaging, and a projection head maps it to 4096 prototypes, where a student matches a teacher that is an exponential moving average of the student itself. We keep the DINO defaults that make this stable, including the

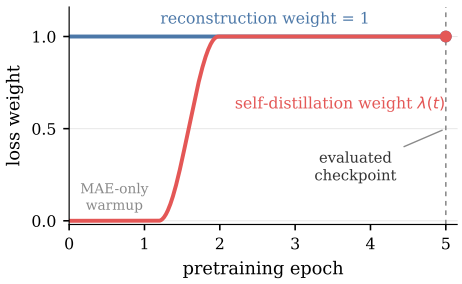

Fig. 5. Objective schedule over the five-epoch run. The reconstruction weight is held at 1 throughout, while the self-distillation weight $\lambda(t)$ stays at 0 for a 15,000-step warmup and then ramps to 1 by roughly epoch 2. The self-distillation branch is thus active for most of pretraining, so the weaker frozen-embedding tasks reflect the short total length rather than a late start.

teacher momentum and temperature ramps and the output centring that stops the embedding collapsing to a constant [1,12].

The two objectives are combined on a schedule rather than from the first step, as Figure 5 shows. We train with reconstruction alone for the first 15,000 steps, then ramp the self-distillation weight $\lambda(t)$ in Equation 1 from 0 to 1 with a cosine curve over the next 10,000 steps, after which it is held at 1. The warmup is needed because the self-distillation teacher is a copy of the encoder. Only after reconstruction has produced a usable feature space do the teacher's targets carry useful signal for the student to match. iBOT and DINOv2 combine a masked-modelling loss with a self-distillation loss on one encoder in the same way [18,12], and we use a 3D convolutional backbone in place of their vision transformer.

3.5 Adaptation to the downstream tasks

We adapt the one pretrained encoder to each task with the smallest change that fits the task type. For the two segmentation tasks we attach the pretrained U-Net decoder and train the full network with an equally weighted sum of Dice and cross-entropy loss. For the three image-level tasks we attach a global-average-pooling head with a linear layer and train with cross-entropy for classification or mean-squared error for regression. When a task provides several co-registered sequences we stack them as input channels, initialising the new stem by repeating the pretrained weights and dividing by the channel count. Finetuning uses AdamW at a learning rate of 10^{-3} and batch size 2, for 50 epochs on the image-level tasks and 150 on the segmentation tasks at a 160^3 patch size, keeping the framework defaults with no per-task hyperparameter search.

4 Experiments

4.1 Setup

Pretraining data. We pretrain on FOMO300K, the method-track corpus of clinical and research brain-MRI scans released with the FOMO26 challenge [2]. After the challenge split we train on 302,741 scans and hold out 3,058 for validation. The corpus pools 36 source collections, the largest contributing 45.9% of the training scans. Grouping by the sequence named in each filename gives 38.7% diffusion-weighted over all b-values, 31.5% T1-weighted, 11.1% T2-weighted, 7.2% FLAIR, 4.3% post-contrast T1, and 3.7% gradient-echo or susceptibility-weighted, with the rest proton-density, perfusion, angiographic, and functional series. Those names are not a usable input specification, because the sources spell them with 337 distinct suffixes and a session holds an arbitrary subset of contrasts, which is the setting the single-channel encoder is built for. Scans are reoriented to RAS and z-score normalised at load time, with no skull stripping, resampling, or bias-field correction, so the encoder sees scans close to their native clinical form at whatever spacing each source provides.

Pretraining configuration. We optimise with AdamW [9] at a learning rate of 10^{-4} and weight decay 3×10^{-5}, with one warmup epoch followed by a cosine decay. Training uses mixed precision on eight A100 GPUs at a global batch size of 24 and a patch size of 160^3, and one epoch is a full pass over the corpus that takes about five hours. We evaluate the checkpoint saved after five epochs, roughly a day of training. This is deliberately short, and it tests whether a light pretraining budget already suffices, given the earlier finding that brain-MRI foundation models saturate early [11]. The weaker frozen-embedding tasks in Section 4.2 show one place where it does not, and we return to the trade-off in Section 4.4.

Downstream tasks and protocol. FOMO26 evaluates one pretrained checkpoint on seven tasks under a few-shot, out-of-domain protocol, adapting on tens of cases from India or the USA and testing on scans from Denmark. The three image-level tasks are scored by AUROC or by mean absolute error and correlation, the two segmentation tasks by Dice and normalised surface distance, and the two no-finetuning tasks by linear-probe separability and group fairness. We report the method-track leaderboard results for the shared checkpoint.

4.2 Results

Table 1 reports the shared checkpoint on the FOMO26 method-track validation leaderboard, where each task drew submissions from 11 or 12 teams. Meningioma segmentation reaches Dice 0.146, polymicrogyria classification AUROC 0.602, and infarct classification AUROC 0.704, all from the five-epoch checkpoint with default finetuning and no hyperparameter search. The absolute values matter

Table 1. The shared checkpoint on the FOMO26 method-track validation leaderboard, over a field of 11 to 12 teams per task. Segmentation is weighted 25% per task and image-level tasks 10%. Task 4 was not scored and is omitted.

Task (metric)	Value
1 Infarct, AUROC	0.704
2 Meningioma, Dice	0.146
3 Brain age, MAE (yr)	14.13
correlation	0.115
5 Polymicrogyria, AUROC	0.602
6 Linear probe, AUROC	0.575
7 Fairness, AUROC	0.949

Table 2. Frozen-probe comparison of the pure masked-autoencoder checkpoint (epoch 0, $\lambda = 0$ throughout its training) against the hybrid checkpoint (epoch 5, $\lambda = 1$). The epoch-0 arm has also seen less data, so the two factors are not separated.

Cohort	n	MAE	Hyb.	Δ
Brain age, R^2	494	0.681	0.646	−0.035
ABIDE age, R^2	120	0.364	0.401	+0.037
ISLES burden, R^2	250	0.080	0.124	+0.044
Tumour, AUROC	29	0.274	0.568	+0.294
Mening., AUROC	45	0.736	0.784	+0.048
Eff. rank, age		185	33	
Eff. rank, mening.		36	10	

more than the positions, and they are modest: few-shot out-of-domain meningioma segmentation is hard for every entry, and against field bests of 0.224 Dice, 0.777 AUROC on infarct, and 9.68 years on brain age, our entries sit mid-field. The frozen embedding gives a linear-probe AUROC of 0.575 at a fairness score of 0.949, where the field is tightly bunched between 0.83 and 0.96. We report the six tasks scored for this checkpoint, omitting trigeminal-nerve segmentation.

4.3 Separating the two objectives

The schedule in Section 3.4 makes one arm of an objective ablation available without a second pretraining run. Because $\lambda(t)$ is held at zero for the first 15,000 steps while one epoch over the corpus is 12,614 steps, the checkpoint saved at the end of epoch 0 has never received a self-distillation gradient, so it is a pure masked-autoencoder model trained by $\mathcal{L}_{\mathrm{MAE}}$ alone. Later checkpoints carry both objectives, since λ reaches 1 during epoch 2. We compare the epoch-0 and epoch-5 checkpoints as frozen feature extractors on five cohorts held out from the challenge test sets and, apart from the in-domain brain-age set, from the pretraining corpus, scoring each by a linear probe on the globally pooled bottleneck.

Table 2 reports the comparison. Adding self-distillation improves four of the five probes, most on the case-control cohorts, where tumour-versus-control separability rises from a below-chance 0.274 to 0.568 AUROC and meningioma-versus-control from 0.736 to 0.784. The one metric favouring the pure-MAE checkpoint is in-domain brain age, the least trustworthy of the five. Effective rank falls from 185 to 33 on brain age and 36 to 10 on meningioma, so self-distillation concentrates the embedding into fewer directions while making it more separable.

Two limits follow from reusing a schedule instead of a controlled pair. The epoch-0 checkpoint has seen one epoch of data and the epoch-5 checkpoint five, so training length and the presence of $\mathcal{L}_{\mathrm{DINO}}$ move together, and we have not

run a self-distillation-only arm. The comparison therefore supports the intended division of labour rather than isolating it. The mask ratio, warmup length, single-channel input, and reconstruction loss follow the reasoning in Sections 3.3 and 3.4, and sweeping each is future work.

4.4 Discussion

The task-level results are consistent with the design and the short pretraining budget. Every number and figure in this paper comes from the same epoch-5 checkpoint. The stronger tasks depend on dense features, which the reconstruction branch builds from the first step, whereas the frozen-embedding and brain-age tasks depend on a linearly separable global embedding that forms more slowly and is normally trained for hundreds of epochs [1,12]. A five-epoch checkpoint is therefore strong where reconstruction dominates and weaker where the embedding does, which matches what we observe, and a longer schedule is the natural next step.

The few-shot protocol also qualifies these numbers, since it adapts on tens of cases, so scores carry substantial variance [11]. Repeating each task's finetuning over three seeds, with the encoder and every hyperparameter fixed, gives a median standard deviation of 0.16 balanced recall on infarct, 0.08 on polymicrogyria, 0.31 years on brain age, and 0.06 Dice on trigeminal segmentation, while the spread between checkpoints is only three to seven times that. Averaging over seeds is the cheap remedy: a three-seed ensemble of the same five-epoch checkpoint raised every task we tried, 0.704 to 0.713 AUROC on infarct, 0.602 to 0.638 on polymicrogyria, 0.575 to 0.589 on the linear probe, and 0.949 to 0.986 on fairness. Each margin is small next to the seed deviation, but the direction is consistent across all four tasks.

5 Conclusion

We presented a 3D brain-MRI foundation model pairing masked reconstruction with self-distillation in a single encoder, so one checkpoint supplies both the dense features segmentation needs and the coherent embedding frozen classification needs, a combination no prior challenge submission applied in a 3D convolutional encoder. On the seven FOMO26 tasks it places mid-field across the scored tasks after five epochs, and the pure masked-autoencoder checkpoint left on the schedule shows self-distillation improving four of five frozen probes. A longer schedule and a controlled pair of runs with λ fixed at 0 and at 1 extend this within the same design. **Disclosure of Interests.** The authors have no competing interests, and this secondary analysis of public, de-identified data required no further review board approval.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
2. Cerri, S., Munk, A., Llambias, S.N., Ambsdorf, J., Machnio, J., et al.: A large-scale heterogeneous 3D magnetic resonance brain imaging dataset for self-supervised learning. arXiv preprint arXiv:2506.14432 (2025)
3. 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 (ICLR) (2021)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
7. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 488–498 (2024)
8. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. In: International Conference on Learning Representations (ICLR) (2022)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
10. Munk, A., Ambsdorf, J., Llambias, S.N., Nielsen, M.: AMAES: Augmented masked autoencoder pretraining on public brain MRI data for 3D-native segmentation. In: MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI (ADSMI) (2024)
11. Munk, A., Cerri, S., Nersesjan, V., Krag, C.H., Ambsdorf, J., et al.: Towards brain MRI foundation models for the clinic: Findings from the FOMO25 challenge. arXiv preprint arXiv:2604.11679 (2026)
12. 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 (TMLR) (2024)
13. Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? In: International Conference on Learning Representations (ICLR) (2023)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
15. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: 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 (CVPR). pp. 20730–20740 (2022)
16. Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Miller, M., Maerkisch, L., Jaeger, P.F., Maier-Hein, K.: Revisiting MAE pre-training for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5186–5196 (2025)

17. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning (ICML). pp. 9929–9939 (2020)
18. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (ICLR) (2022)
19. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021)