This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Materials

Jongjun Won

Neurophet Inc., Seoul, Republic of Korea
wjj910@neurophet.com

S1 Multi-Modal Spark3D Reconstruction

Figure S1 illustrates the self-supervised pretraining task of Sec. 2.2 of the main paper: the network receives a spatially masked, channel-dropped four-sequence patch and reconstructs all four sequences. Only image intensities are used.

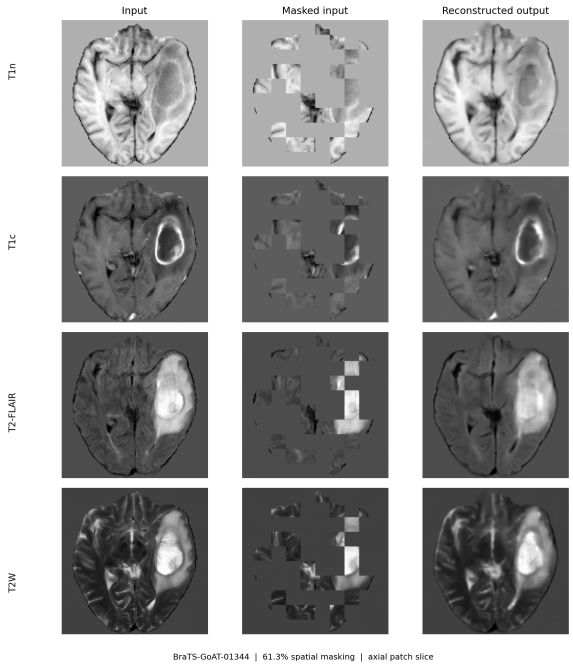


Fig. S1. Representative multi-modal Spark3D reconstruction from a GoAT pretraining patch. After 61.3% spatial masking, the network reconstructs anatomically coherent T1n, T1c, T2-FLAIR and T2W views, while retaining the conspicuous lesion morphology across sequences. No segmentation label is used at any point of pretraining.

S2 Data Usage per Learning Stage

Table S1 states, for every stage of the pipeline, which of the three GoAT splits is used and in which form. The challenge releases 1,351 training cases with expert ground truth, 1,138 further training cases without ground truth, and 451 external validation cases without ground truth, for a total of 2,940 volumes. Two points are worth stating explicitly. First, self-supervised pretraining uses all 2,940 volumes, including the 451 external validation images, as images only. Second, the intermediate pseudo-label training set (1,582 cases) contains 30 external validation cases carrying pseudo-labels generated by our own models, whereas the final training set (2,293 cases) is drawn exclusively from the unlabeled training pool and retained after visual quality control of every pseudo-label. The pseudo-labels for those 942 volumes were produced by the pseudo-label builder, i.e. the ensemble trained on the 1,582-case set.

The 231 cases added in the first pass fall into five groups. Two hundred and one come from the unlabeled training pool: 50 cases carrying the smallest lesions among those whose fold predictions agree, 71 whose fold predictions were empty or near-empty and recover once the decision threshold is lowered, 49 drawn at random from the cases with the highest inter-fold agreement, and 31 whose enhancing tumor was re-assigned inside an unchanged tumor-core mask using the contrast pair – a correction rule verified against ground truth on the labeled training cases before being applied. The remaining 30 are external validation cases to which the last two corrections had been applied. The first three groups target the small and under-detected regions that the exploratory analysis identifies as the dominant error mode.

Table S1. Data usage per learning stage. "Images": volumes used without any segmentation target; "GT": expert ground truth; "PL": pseudo-labels produced by our own models. Numbers in parentheses give how many volumes of that split contribute pseudo-labels to the stage. The hidden test split is never used for training and is omitted.

Learning stage	Labeled training ($n = 1{,}351$)	Unlabeled training ($n = 1{,}138$)	External validation ($n = 451$)
Spark3D SSL pretraining ($n = 2{,}940$)	images	images	images
Supervised fine-tune	images + GT	–	–
Hard-case continued fine-tune	images + GT	–	–
Pseudo-label builder ($n = 1{,}582$)	images + GT	images + PL (201)	images + PL (30)
Modality-dropout dual-view ($n = 2{,}293$)	images + GT	images + PL (942)	–

S3 Stage-to-Stage Change of the Method-Stage Ablation

Table S2 repeats the ablation of Table 1 of the main paper with the change contributed by each stage and with the region-averaged score, which is the quantity the challenge ranks. Reading the columns separately makes the division of labour between the components visible: the enhancing tumor carries most of the total external gain (+0.051 DSC), and almost all of it originates in recall-oriented pseudo-labeling (+0.019) and the modality-dropout dual-base model (+0.018); the tumor core gains +0.026, led by pseudo-labeling (+0.011); and the whole tumor, already the strongest region at the outset, gains +0.019, driven by the dual-base fine-tune (+0.008). Surface distance behaves differently from Dice: the region-averaged NSD stays within 0.509–0.514 across all intermediate stages, and only the final hybrid ensemble improves all three regions simultaneously.

Table S2. Method-stage ablation with stage-to-stage change. Values in parentheses are the change with respect to the preceding row; the *Total* line is the change of the final configuration with respect to the Spark3D baseline. "Mean" is the unweighted average over the three regions. Internal figures are the case-level means of Table 1 of the main paper, whose standard deviations are omitted here for readability; the internal row marked † is optimistically biased by the deliberate training/validation overlap of the low-Dice specialist, so its change should not be read as a contribution.

Configuration	DSC_{ET}	DSC_{TC}	DSC_{WT}	DSC_{Mean}	NSD_{ET}	NSD_{TC}	NSD_{WT}	NSD_{Mean}
External validation – BraTS-GoAT online leaderboard ($n = 451$)								
Spark3D + fine-tune	0.775	0.817	0.877	0.823	0.542	0.498	0.486	0.509
+ low-Dice hard-case FT	0.777 (+0.002)	0.823 (+0.006)	0.881 (+0.004)	0.827 (+0.004)	0.547 (+0.005)	0.508 (+0.010)	0.487 (+0.001)	0.514 (+0.005)
+ pseudo-label retraining	0.796 (+0.019)	0.834 (+0.011)	0.882 (+0.001)	0.837 (+0.010)	0.552 (+0.005)	0.508 (0.000)	0.469 (−0.018)	0.510 (−0.004)
+ modality-dropout dual-base	0.814 (+0.018)	0.832 (−0.002)	0.885 (+0.003)	0.844 (+0.006)	0.567 (+0.015)	0.509 (+0.001)	0.457 (−0.012)	0.511 (+0.001)
+ dual-base fine-tune	0.813 (−0.001)	0.836 (+0.004)	0.893 (+0.008)	0.847 (+0.004)	0.557 (−0.010)	0.505 (−0.004)	0.464 (+0.007)	0.509 (−0.002)
+ hybrid ensemble (final)	**0.826** (+0.013)	**0.843** (+0.007)	**0.896** (+0.003)	**0.855** (+0.008)	**0.566** (+0.009)	**0.507** (+0.002)	**0.482** (+0.018)	**0.518** (+0.010)
Total (final − baseline)	+0.051	+0.026	+0.019	+0.032	+0.024	+0.009	−0.004	+0.010
Internal validation – five-fold CV over the labeled cases ($n = 1{,}351$), case-level means								
Spark3D + fine-tune	0.861	0.909	0.923	0.898	0.885	0.865	0.852	0.867
+ low-Dice hard-case FT†	0.882 (+0.021)	0.932 (+0.023)	0.936 (+0.013)	0.917 (+0.019)	0.910 (+0.025)	0.893 (+0.028)	0.874 (+0.022)	0.892 (+0.025)
+ pseudo-label retraining	0.868 (−0.014)	0.910 (−0.022)	0.925 (−0.011)	0.901 (−0.016)	0.892 (−0.018)	0.864 (−0.029)	0.851 (−0.023)	0.869 (−0.023)
+ modality-dropout dual-base	0.871 (+0.003)	0.912 (+0.002)	0.926 (+0.001)	0.903 (+0.002)	0.896 (+0.004)	0.868 (+0.004)	0.852 (+0.001)	0.872 (+0.003)
+ dual-base fine-tune	0.871 (0.000)	0.915 (+0.003)	0.927 (+0.001)	0.904 (+0.001)	0.896 (0.000)	0.871 (+0.003)	0.852 (0.000)	0.873 (+0.001)

S4 Evidence for the Two Non-Ablated Design Choices

This section collects the published figures behind the two design choices that our own ablation does not isolate, i.e. limitations (i) and (ii) listed in Sec. 4 of the main paper.

S4.1 Contribution of self-supervised pretraining

Every configuration in Table 1 of the main paper starts from the Spark3D-initialized network, so the table quantifies what each *subsequent* stage adds but not what pretraining itself contributes. We did not run the corresponding from-scratch experiment ourselves, but published benchmarks that do apply this control point consistently in one direction. On eleven brain-MRI segmentation tasks, Spark3D improves a from-scratch baseline that shares its architecture, preprocessing and downstream schedule by +1.97 DSC and +1.50 NSD on average (mean rank 2.00 versus 4.55) [2]. The OpenMind benchmark reports the same direction for a ResEnc-L encoder on brain MRI: mean DSC 68.44 from scratch versus 70.36–70.92 with reconstruction-based pretraining [3].

S4.2 Why a CNN backbone

Our choice of the ResEnc-L residual encoder U-Net over a transformer encoder is deliberate. In the most comprehensive benchmark of 3D segmentation architectures to date [1], transformer models (SwinUNETR, SwinUNETRv2, nnFormer, CoTr) fail to match CNN U-Nets on all six datasets evaluated; on BraTS the best transformer reaches DSC 0.9074 against 0.9113–0.9150 for the ResEnc and ConvNeXt variants, and does not surpass the original 2018 nnU-Net (0.9124). The ordering is unchanged in the most recent ConvNeXt scaling study [4], where UNETR and SwinUNETR trail ResEnc-L and MedNeXt by 2.5–6.4 DSC points, and which further reports that from-scratch backbone strength reliably predicts performance after large-scale pretraining – so a backbone that loses from scratch is unlikely to win after self-supervised initialization.

The gap matters most in our regime. Pretraining both a ResEnc-L CNN and a Primus-M transformer on 114k brain-MRI volumes, Wald et al. [3] still find the CNN best overall (mean DSC 70.92 versus 70.42), the transformer 4.8 DSC points weaker from scratch, and its performance collapsing under non-reconstructive pretraining objectives (57.6–63.1 DSC) where the CNN degrades only mildly. A transformer encoder therefore depends on a far larger pretraining corpus than is available here: GoAT permits no external data, so our entire SSL pool is 2,940 volumes – roughly 13× smaller than the 39k used by Spark3D [2] and 39× smaller than OpenMind. Spark3D itself identifies transformer backbones as a pitfall for 3D medical self-supervised learning and adopts the ResEnc U-Net for exactly this reason [2]; its sparse masked-convolution formulation is moreover CNN-specific, so substituting a transformer would replace the pretraining method rather than isolate the backbone. Validating our components on a transformer or ConvNeXt backbone, ideally with a pretraining corpus at the 10k–100k scale, is left to future work.

References

1. 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: MICCAI (2024). arXiv:2404.09556
2. 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: CVPR (2025). arXiv:2410.23132
3. Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Köhler, G., Maier-Hein, K.H.: An OpenMind for 3D medical vision self-supervised learning. arXiv:2412.17041 (2024)
4. Roy, S., Kirchhoff, Y., Ulrich, C., Rokuss, M., Wald, T., Isensee, F., Maier-Hein, K.: MedNeXt-v2: scaling 3D ConvNeXts for large-scale supervised representation learning in medical image segmentation. arXiv:2512.17774 (2025)