Advancing Medical Image Segmentation with Self-Supervised Learning: A 3D Student-Teacher Approach for Cardiac and Neurological Imaging

Moona Mazher[*,1] M.MAZHER@UCL.AC.UK
Daniel C. Alexander[1] D.ALEXANDER@UCL.AC.UK
Abdul Qayyum[2,3] A.QAYYUM@IMPERIAL.AC.UK
Steven A Niederer[2,3] S.NIEDERER@IMPERIAL.AC.UK

[1] *Hawkes Institute, Department of Computer Science, University College London, London, United Kingdom*

[2] *National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, United Kingdom*

[3] *The Alan Turing Institute, London, United Kingdom*

Editors: Under Review for MIDL 2025

Abstract

We propose 3D-SegSync, a novel self-supervised learning (SSL) framework designed to improve segmentation accuracy for both cardiac and neurological structures. 3D-SegSync combines the state-of-the-art DINOv2 student-teacher model architecture with a 3D Vision-LSTM (xLSTM) backbone, which excels at capturing spatiotemporal dependencies and complex anatomical patterns. The SSL phase leverages large-scale unlabeled datasets to pre-train the model, while fine-tuning on labeled data ensures excellent performance across multiple imaging modalities, including CT and MRI. Our framework achieves state-of-the-art results in cardiac and brain image segmentation. 3D-SegSync sets a new benchmark for robust, modality-agnostic medical image segmentation. Code can be found here: https://github.com/Moona-Mazher/3D-SegSync_SSL.

Keywords: Self-Supervised Learning (SSL), Whole Heart Segmentation (WHS), Ischemic Stroke Lesion Segmentation (ISLES), CT Imaging, MRI Imaging, Cardiac Imaging, Neurological Imaging, xLSTM, Multi-Modal Imaging, Traumatic Brain Injury (TBI).

1. Introduction

Medical image segmentation is critical for accurate diagnosis, treatment planning, and monitoring disease progression, especially in complex 3D tasks such as cardiac and neurological imaging. However, segmentation in these areas remains challenging due to factors like limited annotated data, modality variability, and suboptimal image quality. These difficulties are particularly evident in cardiac and brain imaging, where anatomical complexity, patient variability, and motion artifacts add complexity.

Challenges in Cardiac and Neurological Imaging: In cardiac imaging, accurate segmentation of structures like ventricles, atria, myocardium, and blood vessels is essential for diagnosing heart disease. However, the dynamic shape changes across the cardiac cycle,

* Corresponding Author

modality variability (CT vs. MRI), and motion artifacts make segmentation difficult. Existing methods, such as those by (Zhuang and Shen, 2016) and (Isensee et al., 2019), often struggle with multi-center datasets and modality generalization.

In neurological imaging, accurate segmentation of ischemic stroke lesions is critical for prognosis and treatment. While MRI is commonly used, the variability in brain anatomy, lesion complexity, and imaging artifacts present substantial challenges. Models like (Menze et al., 2015) and (Kohl et al., 2020) have demonstrated robust performance but tend to rely on large annotated datasets and struggle with generalization across different clinical settings and modalities.

Gaps in Existing Approaches Ineffective 3D Adaptation: (1) SSL methods like SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) have been successful in 2D tasks but fail to capture the long-range spatial dependencies and complex volumetric data of 3D medical images. (2) Modality-Specific Limitations: Many existing models are optimized for specific imaging modalities (CT or MRI) and struggle to generalize across different modalities, leading to reduced performance in multi-modal settings (Ronneberger et al., 2015a); (Zhu et al., 2021). (3) Dependence on Labeled Data: Despite the promise of SSL, most methods still require substantial labeled datasets for fine-tuning, which remains a bottleneck in medical imaging due to the cost and time involved in manual annotation.

Contribution: We propose a two-stage SSL framework combining DINOv2 with a 3D Vision-LSTM (xLSTM) backbone to capture spatiotemporal dependencies in 3D medical datasets. The model is pretrained on large unlabeled datasets and fine-tuned on smaller labeled datasets for specific segmentation tasks.

Innovations: We extend DINOv2 for 3D medical images using the xLSTM backbone, enabling better capture of spatial dependencies and anatomical features. Our model generalizes across CT and MRI modalities, reducing reliance on labeled data and addressing modality-specific biases.

Overall, Our SSL-based framework addresses the critical limitations of existing methods in both cardiac and neurological imaging. By enhancing robustness, generalizability, and efficiency, our approach sets a new benchmark in medical image segmentation, offering a solution to the challenges of limited labeled data and cross-modality segmentation. This work has the potential to improve clinical decision-making and patient outcomes.

2. Proposed Method

2.1. Dataset

Our study utilised a diverse range of datasets to develop and evaluate our proposed model. For self-supervised learning (SSL) in whole heart segmentation, we used CT Coronary Angiography (CTCA) (Gharleghi et al., 2022) images from the Coronary Atlas, ImageCAS (1,000 patients) (Zeng et al., 2023), ImageTBAD (56 CT angiography images for Type-B aortic dissection segmentation) (Radl et al., 2022), and the TotalSegmentator dataset (1,204 CT scans) (Wasserthal et al., 2023). Additionally, the validation datasets from the "Evaluation of Algorithms for Multi-Modality Whole Heart Segmentation" (MMWHS) challenge (Zhuang et al., 2019) and the "Whole Heart Segmentation++" (WHS++) challenge (Zhuang and Shen, 2016) were incorporated during SSL, while the training samples from MMWHS and WHS++ were used for downstream segmentation tasks. Along with the

MMWHS ans WHS++ we used HVSMR-2.0 (Pace et al., 2024) dataset for the whole heart segmentation task. For neurological/brain imaging tasks, we leveraged ISLES datasets, including ISLES 2022 (de la Rosa et al., 2024) (400 MRI cases) and previous versions (ISLES 2015, 2016, 2018), alongside the ATLAS dataset (Liew et al., 2022), which evolved from 304 cases in v1.2 to 1,271 cases in v2.0. All datasets were incorporated into the SSL framework, enabling the model to learn generalizable feature representations. The model was fine-tuned on ISLES 2024 for stroke lesion segmentation and further validated on Traumatic Brain Injury (TBI) segmentation, demonstrating its robustness across diverse medical imaging tasks.

2.2. Proposed Framework for Heart and Brain Segmentation

Figure 1 presents the overall workflow of the proposed model for whole heart and brain lesion segmentation. The framework comprises four primary stages:

2.2.1. Data Collection and Preprocessing

A variety of datasets, including whole heart CT and MRI scans as well as brain imaging datasets for stroke lesion and TBI segmentation, were curated and preprocessed. This stage involved data augmentation techniques such as flipping, scaling, and noise addition, alongside brightness and contrast adjustments. These transformations were applied to prepare the data for SSL and downstream tasks.

2.2.2. Introduction of a 3D Student-Teacher Model

A 3D student-teacher model, inspired by the 2D DINOv2 (Oquab et al., 2023) framework, is designed on the xLSTM-UNet architecture for the SSL phase. In this model, the teacher encoder's parameters are updated through a momentum-based mechanism derived from the student encoder's updates, ensuring progressive improvements in feature representations. The student encoder is optimized via backpropagation, while the teacher encoder is updated iteratively for consistent learning. This combination of self-supervised learning, momentum updates, and data augmentation enhances the model's ability to deliver accurate results in 3D medical image segmentation. Detailed methodology and mathematical explanations are provided in AppendixC.

2.2.3. Supervised Fine-Tuning for Segmentation

In the third stage, the model was fine-tuned using a limited amount of labelled data for segmentation tasks, including whole heart and stroke lesion segmentation. During this phase, the encoder and bottom layers are fine-tuned to optimise segmentation performance for specific applications.

2.2.4. Evaluation and Performance Analysis

The final stage involved rigorous evaluation of the model's performance through comprehensive analysis. The results were benchmarked against state-of-the-art (SOTA) models (3D-nnUNet (Isensee et al., 2021), 3D-UNet (Ronneberger et al., 2015b), 3D-ResUNet (Li

et al., 2023)), demonstrating the superior accuracy and robustness of the proposed approach in both heart and brain segmentation tasks.

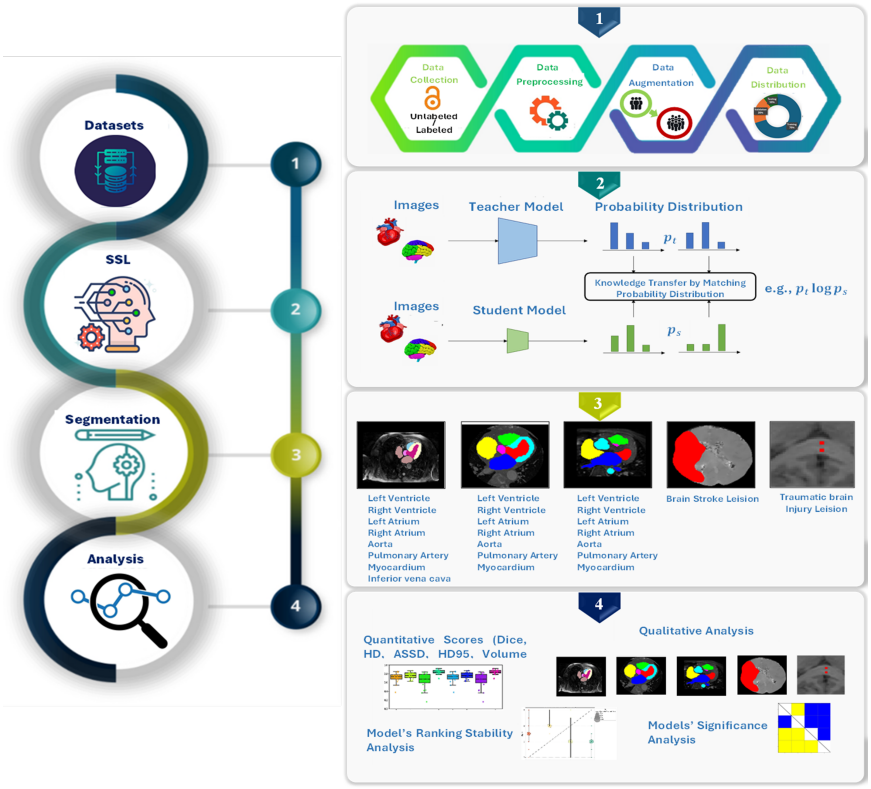

Figure 1: 3D-SegSync framework architecture.

3. Results

We evaluated the performance of our proposed 3D SegSync model on multiple datasets, including three whole-heart (MMWHS, WHS++, HVSMR-2.0 (Pace et al., 2024)), and two neurological/brain imaging (ISLES-2024 stroke and TBI), using an 80/20 training-validation split. Results from Figure 2 show that 3D-SegSync consistently outperforms state-of-the-art (SOTA) models, achieving higher Dice scores and lower Hausdorff Distance 95% (HD95) values across all cardiac imaging datasets, including CT and MRI. In the Dice Coefficient plots (top row), 3D-SegSync demonstrates superior segmentation accuracy with higher medians and a narrow interquartile range (IQR), indicating consistent performance. Similarly, in the HD-95 plots (bottom row), 3D-SegSync achieves lower median values, reflecting reduced segmentation error and greater reliability compared to models like 3D-UNet and 3D-ResNet, which exhibit higher errors and variability.

3D-SegSync, leveraging multi-layer SSL pre-training, achieves significantly higher Dice and lower HD95 scores than its variant 3D-SegSync_bottom, which uses only bottom-layer SSL features. This multi-layer approach captures richer hierarchical representations, en-

abling superior segmentation performance with greater accuracy, lower error rates, and enhanced stability.

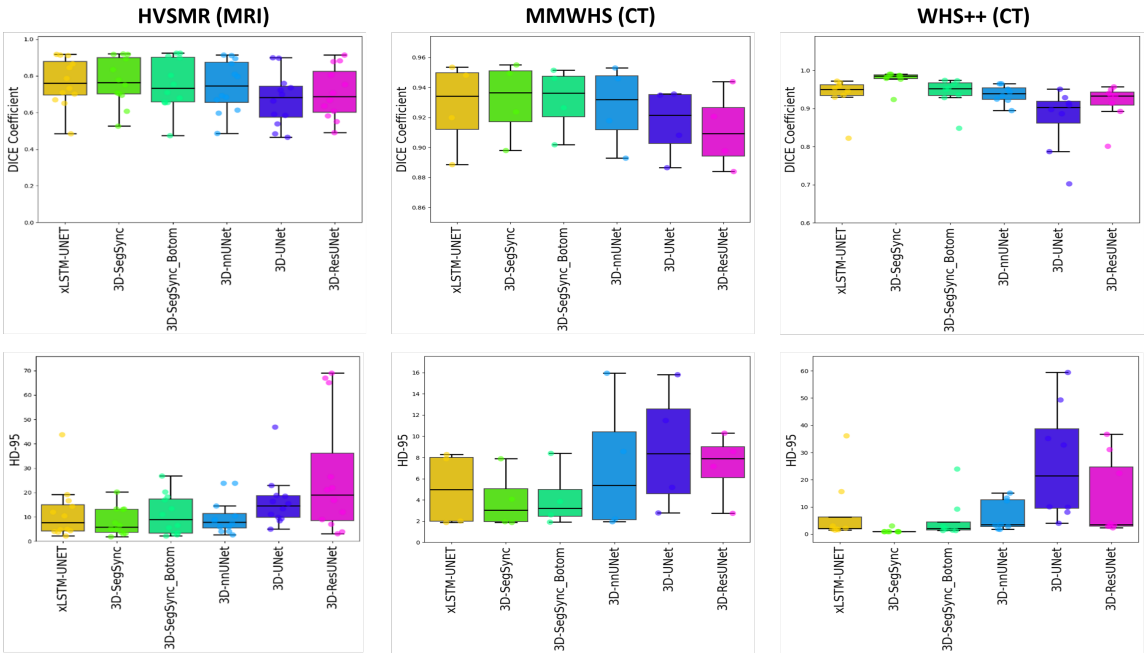

Figure 2: Performance comparison of the proposed 3D-SegSync and SOTA models on Dice and HD-95 metrics across Whole Heart segmentation datasets.

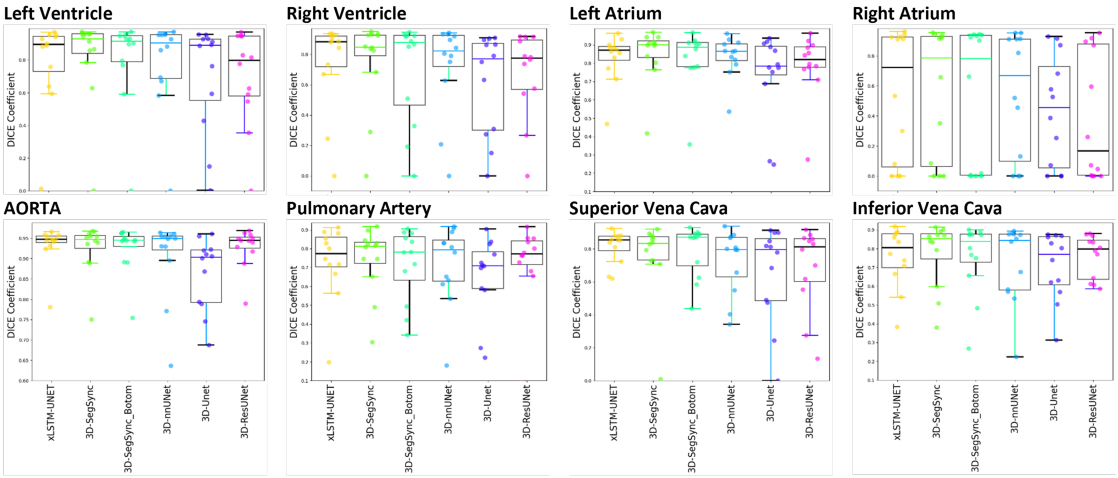

Figure 3: Dice coefficient per label for each model for performance analysis of the proposed model with SOTA approaches on the HVSMR dataset. The labels include LV, RV, and other anatomical structures.

In Figure 3, we presented detailed segmentation results on the HVSMR dataset to showcase 3D-SegSync's performance across all labels. The model excelled in segmenting

anatomical structures like the left ventricle (LV), aorta (AO), and pulmonary artery (PA), achieving superior Dice scores even for smaller or challenging structures. Box plots revealed 3D-SegSync's higher median Dice Coefficient, narrow interquartile range (IQR), and consistent performance, even with outliers, highlighting its robustness. This improvement stems from advanced multi-layer SSL pre-training, enabling richer feature representation. Unlike 3D-SegSync-bottom, the full model integrates high-level context for better segmentation accuracy and fine boundary capture, as evidenced by lower HD95 values. Further analysis of generalization across imaging modalities and significance maps is provided in AppendixB.

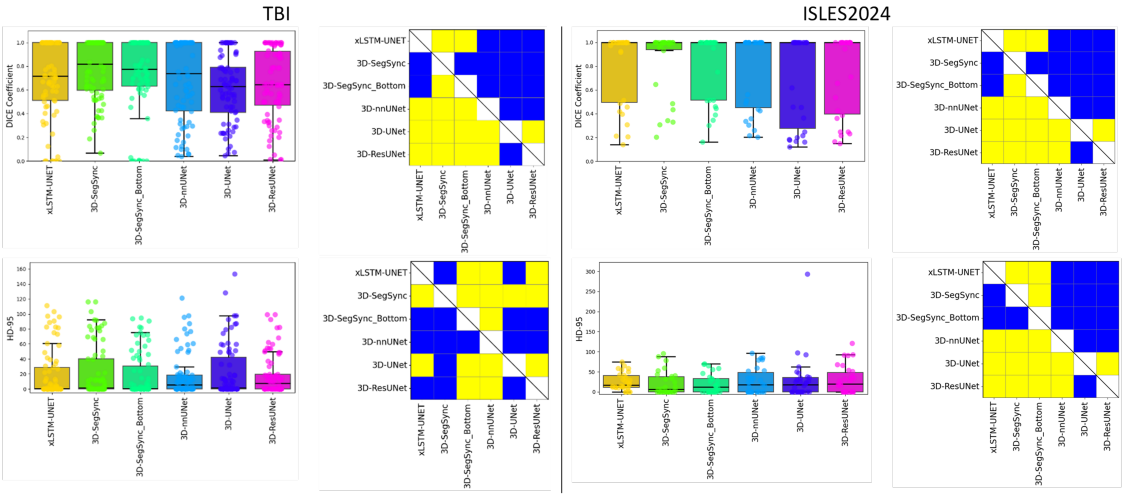

Figure 4: Performance comparison of the proposed 3D-SegSync and SOTA models on Dice and HD-95 metrics across TBI and ISLES brain lesion segmentation datasets.

Table 1: Performance analysis of 3D-SegSync with its variants and SOTA models for all heart and brain imaging datasets.

Average DICE Coefficient						
	3D-SegSync	**3D-SegSync_Botom**	**xLSTM-UNET**	**3D-nnUNet**	**3D-ResUNet**	**3D-UNet**
HVSMR-2.0	0.77132	0.753696	0.766224	0.745084	0.706659	0.670413
MMWHS CT	0.941531	0.931573	0.927654	0.917473	0.891598	0.886358
MMWHS MRI	0.87167	0.86413	0.86338	0.85904	0.84187	0.83663
WHS++ CT	0.976996	0.941146	0.935651	0.927362	0.915044	0.872164
WHS++ MRI	0.887617	0.886642	0.87293	0.869232	0.851003	0.858753
TBI	0.782532	0.724659	0.686121	0.678905	0.678905	0.643323
ISLES2024	0.848296	0.804858	0.798529	0.769919	0.742434	0.702374
Average HD 95						
HVSMR-2.0	17.25634	24.21755	29.22956	22.12653	68.07598	35.08616
MMWHS CT	14.39735	18.868	17.65168	19.64561	35.8839	58.16591
MMWHS MRI	29.02241	46.32272	51.50929	23.19734	42.48808	86.60793
WHS++ CT	5.281255	17.94436	21.58496	29.60166	64.24364	80.54156
WHS++ MRI	12.19138	13.17895	16.08778	16.02272	20.14479	16.35387
TBI	19.45671	23.17893	24.13334	25.31134	27.21234	28.18971
ISLES2024	29.22114	31.67891	34.72802	35.09987	39.61133	38.87633

Figure 4 compares the performance of 3D-SegSync with state-of-the-art models (3D-SegSync-bottom, xLSTM-UNet, 3D-nnUNet, 3D-ResUNet, 3D-UNet) on the ISLES2024

stroke and TBI datasets. 3D-SegSync outperforms all models, achieving the highest Dice scores and lowest HD95 values.

Table 1 presents the Dice and HD95 scores across heart and brain imaging datasets, demonstrating that 3D-SegSync outperforms all models, with 3D-SegSync-bottom coming second. This variant fine-tunes only the bottom-layer features during segmentation training. The results indicate that optimizing all layers of the encoder leads to better performance. A detailed explanation of each dataset can be found in AppendixB.

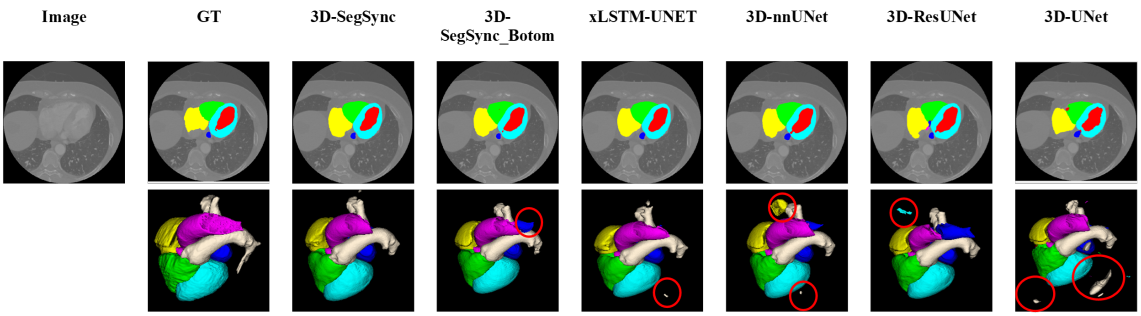

Figure 5: Quantitative Performance of the Proposed and SOTA Models on the MMWHS CT Dataset. Colour representation: Purple (AO), Yellow (RA), Red (LV), Light Blue (Myo), Gray (PA), Blue (LA), Green (RV).

Figure 5 illustrates the quantitative performance of the proposed 3D-SegSync on the MMWHS_CT dataset. The results demonstrate that the proposed 3D-SegSync model achieves a segmentation output that is closely aligned with the ground truth (GT) segmentation map, outperforming other SOTA models across most anatomical regions of the whole heart. Among the comparative models, 3D-ResUNet and 3D-UNet demonstrate a higher rate of segmentation errors, especially in the pulmonary veins and aorta. The SOTA 3D-nnUNet, while performing comparatively better, exhibits noticeable errors in the right atrium, as shown in the 3D segmentation map. These observations provide valuable insights into the potential areas for further refinement in cardiac segmentation methods.

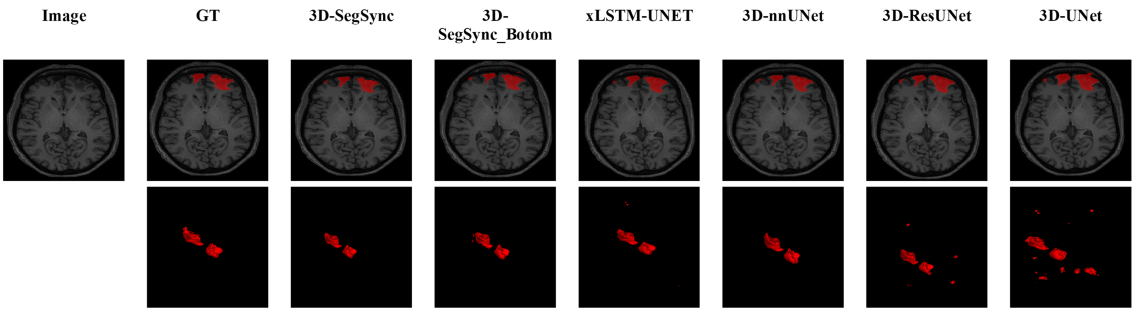

Figure 6: Quantitative analysis of the proposed 3D-SegSync model compared to state-of-the-art (SOTA) models for TBI lesion segmentation.

Beyond whole-heart segmentation, we validated the efficacy of our proposed model on the neurological/brain imaging datasets such as the TBI dataset (see Figure 4) from the

MICCAI (Medical Image Computing and Computer Assisted Intervention) 2024 Grand Challenge. Our model secured first place in the TBI validation and testing phases, demonstrating its exceptional accuracy and generalisability. The leaderboard for the TBI challenge can be accessed at https://aims-tbi.grand-challenge, where our team, DeepLearnAI, is listed in the top position. Additionally, we tested our model on the ISLES-2024 stroke challenge (see Figure 4), achieving first place on the leaderboard under the team name Dolphins. The leaderboard for the ISLES challenge can be viewed at https://isles-24.grand. These achievements on both TBI and ISLES-2024 challenges underline the superior performance of our proposed model compared to other SOTA deep learning approaches.

Figure 6 presents a quantitative analysis of the proposed 3D-SegSync model compared to state-of-the-art (SOTA) models for TBI lesion segmentation. It demonstrates that the 3D-SegSync model outperforms state-of-the-art models in segmentation accuracy, aligning closely with the ground truth. In contrast, the baseline xLSTM model shows significant limitations. The complex and heterogeneous nature of lesions in moderate to severe TBI (msTBI) presents challenges for accurate neuroimaging analysis. 3D-SegSync effectively addresses these challenges by capturing the intricate features of msTBI lesions, enabling precise segmentation even in the presence of high variability.

The success of 3D-SegSync across various datasets and modalities highlights its robustness and versatility. By leveraging pre-trained SSL features, the model reduces reliance on large labeled datasets, making it highly effective in medical imaging where annotated data is limited. These results establish 3D-SegSync as a reliable solution for diverse medical image segmentation challenges.

Future work could involve applying SSL to larger datasets for improved generalizability, extending 3D-SegSync to other imaging modalities (e.g., ultrasound, PET), and incorporating multi-modal data (e.g., clinical or genomic data) to improve diagnostic accuracy. Incorporating interpretability techniques could further enhance trust in clinical applications. Addressing these areas will help 3D-SegSync evolve into a more powerful tool for medical imaging.

4. Conclusion

We introduced 3D-SegSync, a robust 3D medical image segmentation framework designed to address challenges like data scarcity, modality variability, and anatomical complexity. By combining the DINOv2 teacher-student architecture with the xLSTM-UNet, 3D-SegSync leverages self-supervised learning to extract rich, modality-independent 3D features from large-scale unlabelled datasets. The xLSTM-UNet further enhances the model's ability to capture spatial and contextual relationships in 3D imaging, making it highly effective for segmentation tasks. This fully 3D framework achieves state-of-the-art performance in whole-heart, stroke lesion, and traumatic brain injury segmentation across CT and MRI, significantly reducing dependence on labeled datasets. By uniting powerful 3D self-supervised learning with efficient design, 3D-SegSync sets a new benchmark in medical imaging, offering improved scalability and clinical relevance. Future work will explore its application to broader tasks, strengthening its cross-modality capabilities and impact.

Acknowledgments

We would like to express our gratitude to all those who contributed to this work. Their insightful feedback, support, and collaboration were invaluable.

References

Tianrun Chen, Chaotao Ding, Lanyun Zhu, Tao Xu, Deyi Ji, Yan Wang, Ying Zang, and Zejian Li. xlstm-unet can be an effective 2d & 3d medical image segmentation backbone with vision-lstm (vil) better than its mamba counterpart. *arXiv preprint arXiv:2407.01530*, 2024.

X. Chen et al. A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

Ezequiel de la Rosa, Mauricio Reyes, Sook-Lei Liew, Alexandre Hutton, Roland Wiest, Johannes Kaesmacher, Uta Hanning, Arsany Hakim, Richard Zubal, Waldo Valenzuela, et al. A robust ensemble algorithm for ischemic stroke lesion segmentation: Generalizability and clinical utility beyond the isles challenge. *arXiv preprint arXiv:2403.19425*, 2024.

Ramtin Gharleghi, Dona Adikari, Katy Ellenberger, Sze-Yuan Ooi, Chris Ellis, Chung-Ming Chen, Ruochen Gao, Yuting He, Raabid Hussain, Chia-Yen Lee, et al. Automated segmentation of normal and diseased coronary arteries–the asoca challenge. *Computerized Medical Imaging and Graphics*, 97:102049, 2022.

K. He et al. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Fabian Isensee et al. Automated design of deep learning methods for biomedical image segmentation. *Medical Image Analysis*, 58:101559, 2019.

S. K. Kohl et al. A deep learning framework for stroke lesion segmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1277–1287, 2020.

Xiangyu Li, Gongning Luo, Kuanquan Wang, Hongyu Wang, Jun Liu, Xinjie Liang, Jie Jiang, Zhenghao Song, Chunyue Zheng, Haokai Chi, et al. The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: The instance challenge. *arXiv preprint arXiv:2301.03281*, 2023.

Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

Moona Mazher, Imran Razzak, Abdul Qayyum, M Tanveer, Susann Beier, Tariq Khan, and Steven A Niederer. Self-supervised spatial–temporal transformer fusion based federated framework for 4d cardiovascular image segmentation. *Information Fusion*, 106:102256, 2024.

B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2004, 2015.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Danielle F Pace, Hannah TM Contreras, Jennifer Romanowicz, Shruti Ghelani, Imon Rahaman, Yue Zhang, Patricia Gao, Mohammad Imrul Jubair, Tom Yeh, Polina Golland, et al. Hvsmr-2.0: A 3d cardiovascular mr dataset for whole-heart segmentation in congenital heart disease. *Scientific Data*, 11(1):721, 2024.

Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-hua Zhao, and Jan Egger. Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. *Data in brief*, 40:107801, 2022.

O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015a.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015b.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

An Zeng, Chunbiao Wu, Guisen Lin, Wen Xie, Jin Hong, Meiping Huang, Jian Zhuang, Shanshan Bi, Dan Pan, Najeeb Ullah, et al. Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. *Computerized Medical Imaging and Graphics*, 109:102287, 2023.

X. Zhu et al. Deep learning for cardiac image segmentation: Challenges and opportunities. *Medical Image Analysis*, 70:101980, 2021.

Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016.

Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

Appendix A. Proposed 3D-SegSync Architecture

See the Figure 7

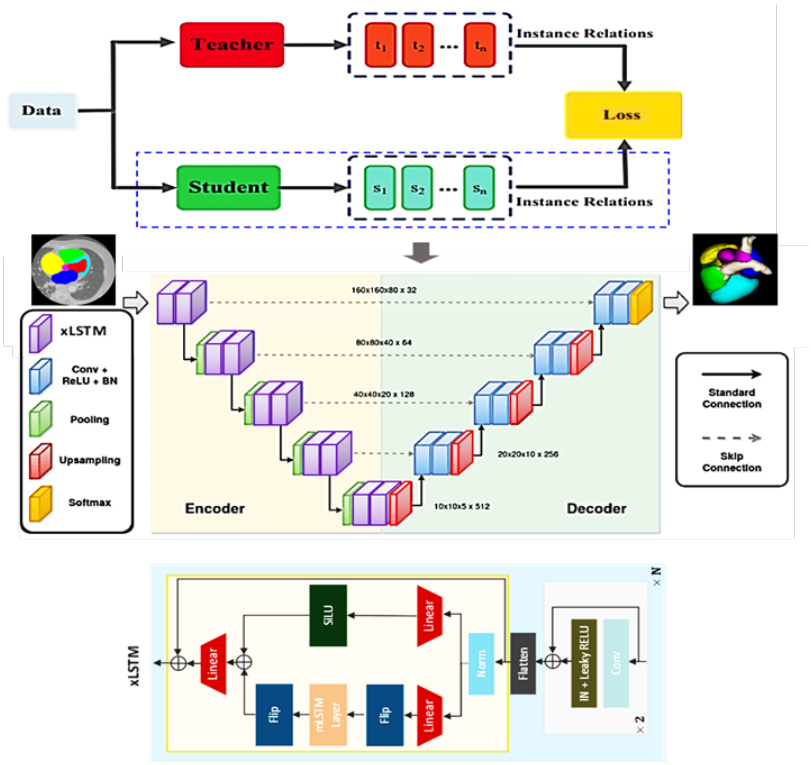

Figure 7: Proposed 3D-SegSync model architecture.

Appendix B. Comparsions of 3D-SegSync with SOTA models

Figure 9 demonstrates 3D-SegSync's ability to generalise across different imaging modalities, out-performing SOTA models in both CT and MRI datasets for whole-heart segmentation. This ability to learn modality-independent features via SSL pre-training ensures its applicability in clinical settings where multimodal imaging is common.

Figure 8 presents significance maps, where 3D-SegSync consistently shows higher yellow regions, indicating statistically significant improvements in Dice and HD95 metrics compared to other models. 3D SegSync_bottom shows fewer yellow regions in comparison with its advance version 3D-SegSync, reflecting its weaker performance, while 3D-UNet displays highest blue regions, indicating significantly lowest performance among all models in most

of the datasets such as HVSMR and WHS++. We have also given the model ranks on the all whole heart segmentation datasets in Figure 10.

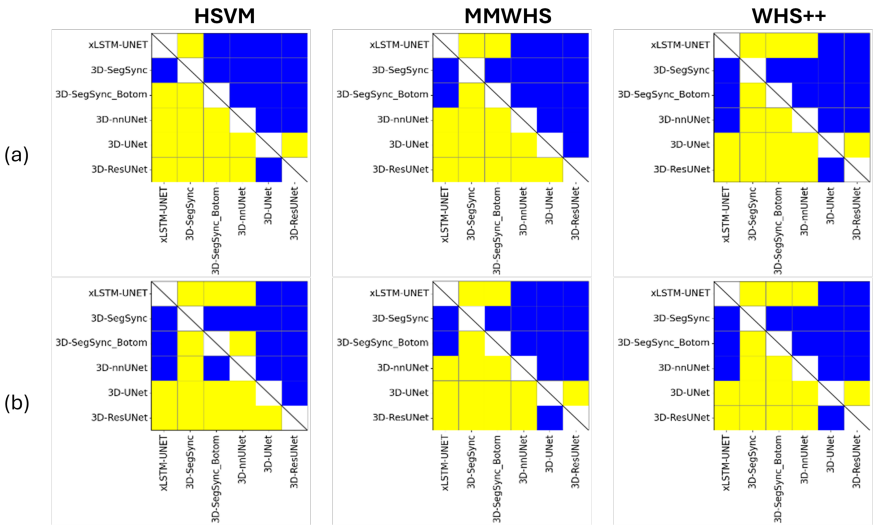

Figure 8: Significance maps of the proposed 3D-SegSync and SOTA models on DICE (a) and HD-95 (b) metrics across Whole Heart segmentation datasets.

Table 2: Performance analysis of proposed and state of the art models for HVSMR-2.0 dataset.

Model	Dice_Avg_All	HD_Avg_All	HD95_Avg_All	ASSD_Avg_All	Vol_Diff_Avg_All
3D-SegSync	0.77132	17.25634	8.006959	2.02056	0.02273048
3D-SegSync_Botom	0.753696	24.21755	10.89928	3.03842	0.024452868
xLSTM-UNET	0.766224	29.22956	11.74866	3.100872	0.026769222
3D-nnUNet	0.745084	22.12653	10.00376	2.39672	0.026893855
3D-ResUNet	0.706659	68.07598	26.91917	7.344615	0.037925875
3D-UNet	0.670413	35.08616	16.39909	4.138591	0.037455937

Table 3: Performance analysis of proposed and SOTA models using MMWHS CT dataset.

Model	Dice_Avg	HD_Avg	HD95_Avg	ASSD_Avg	Vol_Diff_Avg
3D-SegSync	0.941531281	14.3973592	4.197811849	0.942519917	0.006214206
3D-SegSync-Botom	0.93157309	18.8680027	3.95788033	0.964365257	0.006409229
xLSTM-UNET	0.927654749	17.6516858	5.022211073	1.014969103	0.006599423
3D-nnUNet	0.917473901	19.6456178	7.166018194	1.330353153	0.006746526
3D-ResUNet	0.891598521	35.883997	7.205884387	1.666301144	0.006330065
3D-UNet	0.886358506	58.1659145	8.820965598	2.036248398	0.008241135

Table 2 shows the segmentation results on the HVSMR-2.0 whole-heart MRI dataset demonstrate that 3D-SegSync achieved the best overall performance, with the highest Dice score (0.7713), the lowest ASSD (2.0206 mm), and the smallest volume difference (0.0227), indicating accurate overlap, surface alignment, and volume estimation. 3D-SegSync_Botom

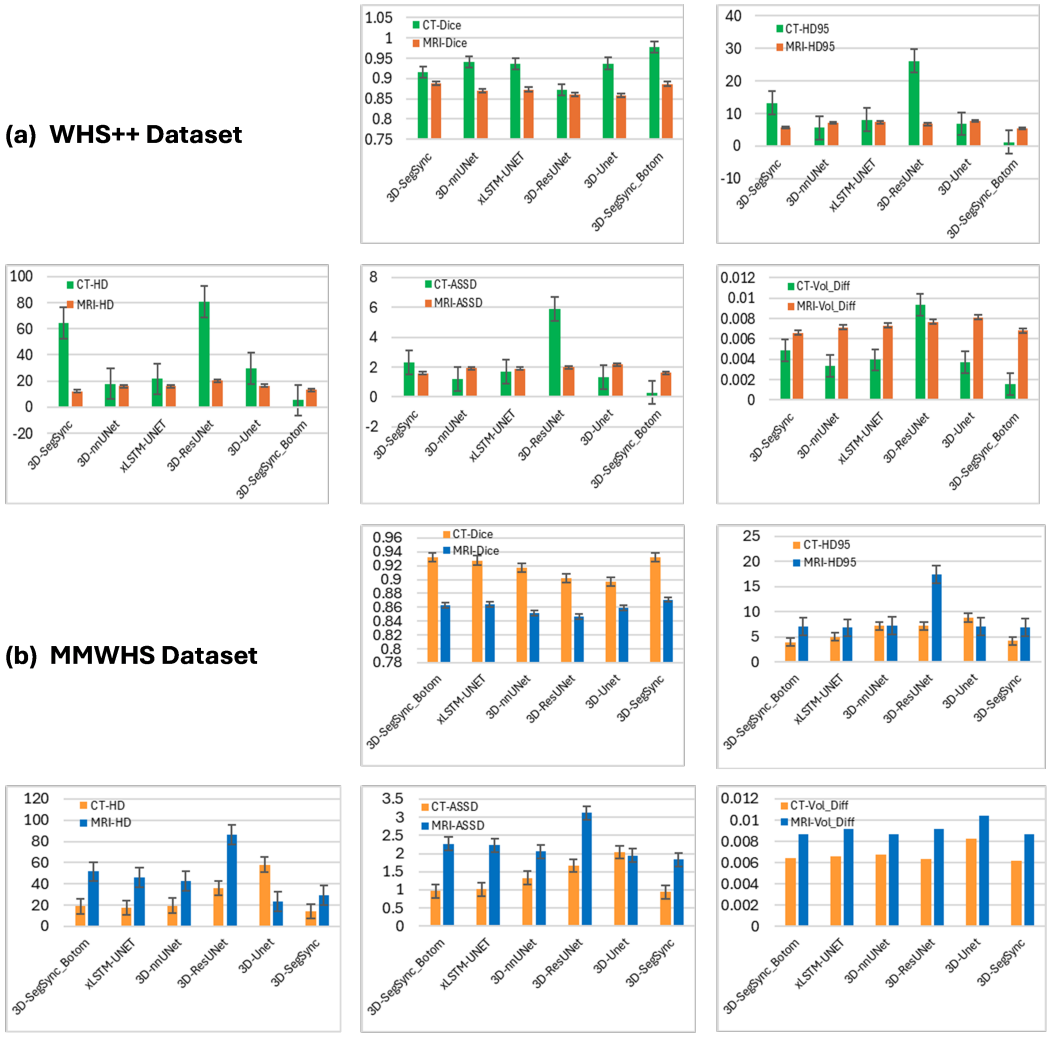

Figure 9: Cross modality performance comparison of 3D-Segsync and SOTA models for whole-heart howsegmentation across CT and MRI datasets. (a) WHS++ dataset, where green bars show CT and mustard bars show MRI. (b) MMWHS dataset, where orange bars show CT and blue bars show MRI.

and xLSTM-UNET also performed well, with Dice scores of 0.7537 and 0.7662, respectively, though both exhibited higher Hausdorff distances (HD95 of 10.8993 mm and 11.7487 mm) and ASSD values, reflecting less precise boundary alignment. While 3D-nnUNet showed good surface alignment (ASSD of 2.3967 mm), its lower Dice score (0.7451) and higher HD95 (10.0038 mm) suggest moderate segmentation accuracy. In contrast, 3D-ResUNet and 3D-UNet underperformed, with significantly lower Dice scores (0.7067 and 0.6704) and much higher HD95 (26.9192 mm and 16.3991 mm), indicating poor boundary and surface alignment. Overall, 3D-SegSync is the most reliable model for whole-heart segmentation in this dataset.

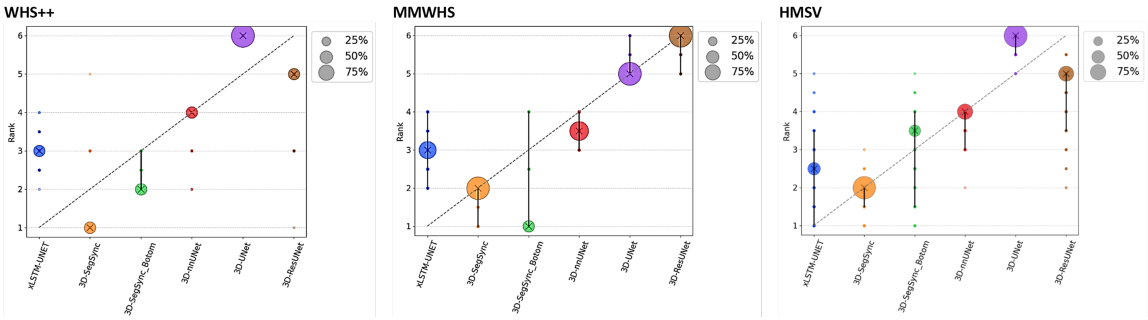

Figure 10: Blob plots illustrating the stability of rankings of whole heart segmentation datasets based on bootstrap sampling. The median rank for each algorithm is represented by a black cross, while the 95% bootstrap intervals across samples are depicted as black lines.

Table 3 shows the segmentation results on the MMWHS dataset demonstrate that 3D-SegSync achieved the best overall performance, with the highest Dice score (0.9415), the lowest ASSD (0.9425 mm), and a minimal volume difference (0.0062), indicating excellent overlap, surface alignment, and volume estimation. 3D-SegSync-Botom also performed well, with a Dice score of 0.9316 and the lowest HD95 (3.9579 mm), though it showed slightly higher ASSD (0.9644 mm). xLSTM-UNET attained a Dice score of 0.9277 but exhibited higher HD95 (5.0222 mm) and ASSD (1.0150 mm), reflecting less precise boundary alignment. 3D-nnUNet demonstrated moderate performance, with a Dice score of 0.9175 and higher HD95 (7.1660 mm) and ASSD (1.3304 mm). In contrast, 3D-ResUNet and 3D-UNet underperformed, with significantly lower Dice scores (0.8916 and 0.8864) and considerably higher HD95 (7.2059 mm and 8.8210 mm) and ASSD (1.6663 mm and 2.0362 mm), indicating poor boundary and surface alignment. Overall, 3D-SegSync is the most effective model for whole-heart segmentation in this dataset.

Tables 4, 5, and 6present the results of CT and MRI whole-heart segmentation, where the proposed 3D-SegSync model, employing a self-supervised learning approach, consistently outperformed state-of-the-art (SOTA) models across Dice and other key performance metrics. The model demonstrated superior accuracy in overlap, boundary alignment, and volume estimation, validating its effectiveness for both CT and MRI modalities. Additionally, the model was evaluated on the WHS++ dataset, where similar performance trends were observed, reinforcing its generalisability to different datasets.

Furthermore, the proposed 3D-SegSync was assessed against SOTA models for stroke lesion segmentation (ISLES2024) and traumatic brain injury (TBI) lesion segmentation tasks. In these evaluations (Tables 7 and 8), 3D-SegSync consistently delivered robust and reliable performance, showcasing its versatility and efficacy in segmenting diverse anatomical and pathological structures. These results highlight the potential of the self-supervised 3D-SegSync model to set a new standard in medical image segmentation across multiple domains.

Table 4: Performance analysis of proposed and SOTA models using MMWHS MRI dataset.

Model	**Dice_Avg**	**HD_Avg**	**HD95_Avg**	**ASSD_Avg**	**Vol_Diff_Avg**
3D-SegSync	0.87167	29.02241	6.871293	1.831784	0.008656418
3D-SegSync_Botom	0.86413	46.32272	6.813768	2.231868	0.009208905
xLSTM-UNET	0.86338	51.50929	7.029732	2.266159	0.008637976
3D-nnUNet	0.85904	23.19734	7.03841	1.946138	0.010388217
3D-ResUNet	0.84187	42.48808	7.26451	2.05212	0.00868741
3D-UNet	0.83663	86.60793	17.44347	3.115857	0.009206529

Table 5: Performance analysis of proposed and SOTA models using WHS++ CT dataset.

Model	**Dice_Avg**	**HD_Avg**	**HD95_Avg**	**ASSD_Avg**	**Vol_Diff_Avg**
3D-SegSync	0.976996	5.281255	1.251489	0.301377	0.001588615
3D-SegSync_Botom	0.941146	17.94436	5.572338	1.184542	0.003344029
xLSTM-UNET	0.935651	21.58496	8.052558	1.677745	0.003928403
3D-nnUNet	0.927362	29.60166	6.834203	1.301429	0.003741077
3D-ResUNet	0.915044	64.24364	13.20491	2.291552	0.004881045
3D-UNet	0.872164	80.54156	26.13814	5.87906	0.00934523

Table 6: Performance analysis of proposed and SOTA models using WHS++ MRI dataset.

Model	**Dice_Avg**	**HD_Avg**	**HD95_Avg**	**ASSD_Avg**	**Vol_Diff_Avg**
3D-SegSync	0.887617	12.19138	5.696797	1.577587	0.006581231
3D-SegSync_Botom	0.886642	13.17895	5.460724	1.625685	0.006799575
xLSTM-UNET	0.87293	16.08778	7.373174	1.892255	0.007296956
3D-nnUNet	0.869232	16.02272	7.136496	1.927073	0.007132139
3D-ResUNet	0.851003	20.14479	6.741368	1.969676	0.007659463
3D-UNet	0.858753	16.35387	7.723141	2.168973	0.008124995

Table 7: Performance analysis of proposed and SOTA models using TBI MRI dataset.

Model	**Dice_Avg**	**HD_Avg**	**HD95_Avg**	**ASSD_Avg**	**Vol_Diff_Avg**
3D-SegSync	0.782532	19.45671	15.36666	2.45671	0.275341
3D-SegSync_Botom	0.724659	23.17893	17.99199	3.01345	0.492328
xLSTM-UNET	0.686121	24.13334	19.92581	3.18965	0.718707
3D-nnUNet	0.678905	25.31134	17.84589	3.30567	0.739339
3D-ResUNet	0.678905	27.21234	17.84589	4.17865	0.786737
3D-UNet	0.643323	28.18971	18.62511	6.34567	0.852835

Table 8: Performance analysis of proposed and SOTA models using ISLES2024 dataset.

Model	Dice_Avg	HD_Avg	HD95_Avg	ASSD_Avg	Vol_Diff_Avg
3D-SegSync	0.848296	29.22114	21.05371	2.78653	0.840097
3D-SegSync_Botom	0.804858	31.67891	20.2839	1.89765	0.840068
xLSTM-UNET	0.798529	34.72802	25.56568	2.93112	0.865685
3D-nnUNet	0.769919	35.09987	27.30891	2.98765	0.887896
3D-ResUNet	0.74243459	39.61133	30.11339	4.23478	0.900569
3D-UNet	0.702374	38.87633	30.9896	6.13459	0.946751

Appendix C. Methodology & Mathematical Details of the Proposed Framework

C.1. Methodology

The proposed framework is built on a self-supervised learning (SSL) (Mazher et al., 2024) approach designed to pre-train a 3D Vision-LSTM (xLSTM) integrated UNet model (xLSTM-UNet) (Oquab et al., 2023; Chen et al., 2024). The methodology combines advanced deep learning techniques to achieve enhanced performance in 3D medical image segmentation tasks. The main diagram of proposed SSL model is shown in AppendixA.

C.1.1. Data Augmentation in the Student-Teacher Framework

Robust data augmentation plays a critical role in the SSL pipeline. Techniques such as flipping, scaling, Gaussian noise addition, Gaussian blur, and adjustments to brightness and contrast are applied to create diverse and informative training inputs. Two augmented views of each input image are generated and processed through a Siamese network structure, comprising the student and teacher encoders. The teacher encoder's outputs are refined through centring, sharpening, and normalisation via a softmax function, producing supervision signals for the student encoder.

The loss function ensures alignment between the student's outputs and the teacher's processed outputs by minimising divergence, employing cross-entropy loss and mean squared error (MSE) (Oquab et al., 2023). This alignment facilitates robust feature learning from unlabelled data, enhancing the model's generalisation capabilities.

C.1.2. xLSTM-UNet Architecture

The xLSTM-UNet model (Chen et al., 2024) integrates Vision-LSTM (xLSTM), an advanced extension of Long Short-Term Memory (LSTM) networks, into the UNet architecture. xLSTM excels at capturing long-range dependencies and contextual information, complementing the UNet's strength in extracting local features through its convolutional encoder-decoder design. The encoder identifies hierarchical features from the input, while the decoder reconstructs these features into detailed segmentation maps, enabling precise and reliable segmentation.

C.1.3. Self-Supervised Pre-Training and Supervised Fine-Tuning

The SSL framework focuses on pre-training the xLSTM-UNet encoder using unlabelled data to capture meaningful spatial and contextual features. Once pre-trained, the encoder is fine-tuned in a supervised manner using labelled datasets, optimising the decoder to generate accurate segmentation maps. This two-stage process minimises the reliance on extensive labelled datasets, while the xLSTM module ensures effective learning of global context and long-range dependencies.

C.2. Mathematical Details of the Proposed Framework

The momentum teacher encoder's parameters θ_t are updated based on the student encoder's parameters θ_s using a momentum-based approach:

$$\theta_t = m \cdot \theta_t + (1 - m) \cdot \theta_s \tag{1}$$

Where θ_t are the parameters of the teach encoder, θ_s are the parameters of the student encoder, m is the momentum coefficient typically a value close to 1.

Let x be the original input image. Two different views of the input, x_1 and, x_2 are generated using strong data augmentations:

$$x_1 = \text{Augment}(x), \quad x_2 = \text{Augment}(x) \tag{2}$$

Both views are then processed through the student encoder f_s and teacher encoder f_t to extract feature representations:

$$h_1 = f_s(x_1; \theta_s), \quad h_2 = f_s(x_2; \theta_s) \tag{3}$$

$$h'_1 = f_s(x_1; \theta_t), \quad h'_2 = f_s(x_2; \theta_t) \tag{4}$$

Where h_1 and h_2 are the feature representations from the student encoder, h'_1 and h'_2 are the feature representations from the teacher encoder.

The feature representations h_1, h_2, h'_1 , h'_2 are subjected to global average pooling to reduce them into feature vectors:

$$v_1 = \text{GAP}(h_1), \quad v_2 = \text{GAP}(h_2) \tag{5}$$

$$v'_1 = \text{GAP}(h'_1), \quad v'_2 = \text{GAP}(h'_2) \tag{6}$$

Where v_1, v_2, v'_1 and v'_2 are the resulting feature vectors.

$$z_1 = \text{MLP}(v_1), \quad z_2 = \text{MLP}(v_2) \tag{7}$$

$$z'_1 = \text{MLP}(v'_1), \quad z'_2 = \text{MLP}(v'_2) \tag{8}$$

After projection, the teacher's output is centered, sharpened, and passed through a softmax function to produce the supervision signal:

$$q_1' = \text{Softmax}\left(\frac{\text{Center}(z_1')}{\tau}\right) \tag{9}$$

$$q_2' = \text{Softmax}\left(\frac{\text{Center}(z_2')}{\tau}\right) \tag{10}$$

Where $Center(z)$ subtracts the mean of the vector to have zero mean. τ is the temperature parameter controlling the sharpness of the distribution. $Softmax(z)$ normalizes the vector into a probability distribution.

The loss function is designed to minimize the divergence between the student's feature vectors and the teacher's processed outputs. A common choice is the cross-entropy loss or mean squared error (MSE) between the student's and teacher's outputs:

$$L = \frac{1}{2}\left(\text{Loss}(z_1, q_2') + \text{Loss}(z_2, q_1')\right) \tag{11}$$

Where this loss function encourages the student encoder to produce feature representations that align closely with the teacher's outputs, thus enabling effective learning from the unlabeled data.

For the downstreaming segmentation training task we used the cross-entropy loss as given in the equation below :

$$L(z, q') = -\sum_{k=1}^{K} q'[k] \log(\text{Softmax}(z)[k]) \tag{12}$$