Continual Self-supervised Learning: Towards Universal Multi-modal Medical Data Representation Learning

Yiwen Ye[1] Yutong Xie[2†] Jianpeng Zhang[3] Ziyang Chen[1] Qi Wu[2] Yong Xia[1,4,5†]

[1] *School of Computer Science and Engineering, Northwestern Polytechnical University, China*

[2] *Australian Institute for Machine Learning (AIML), The University of Adelaide, Australia*

[3] *College of Computer Science and Technology, Zhejiang University, China*

[4] *Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China*

[5] *Ningbo Institute of Northwestern Polytechnical University, China*

ywye@mail.nwpu.edu.cn, yutong.xie678@gmail.com, jianpeng.zhang0@gmail.com

zychen@mail.nwpu.edu.cn, qi.wu01@adelaide.edu.au, yxia@nwpu.edu.cn

A. Architecture Details on Downstream Tasks

A.1. Classification tasks

We utilized a tokenizer chosen from the set $\{\mathcal{T}^{1d}, \mathcal{T}^{2d}, \mathcal{T}^{3d}\}$ based on the dimensionality of input data, the encoder ϕ, and the newly appended Multi-Layer Perceptron (MLP) head to output the desired predictions. The architectures of these MLP heads are tailored to suit the specific requirements of individual datasets.

PudMed20k dataset and RICORD dataset: A two-layer MLP head is adopted as the head. The first layer has 768 neurons followed by a layer normalization layer and the Gaussian error linear units (GELU). The second layer decreases the number of neurons to 5 and 2 for the PudMed20k and RICORD datasets, respectively.

ChestXR dataset and NCH dataset: A one-layer MLP head is adopted as the head. The number of neurons is 3 and 9 on ChestXR and NCH datasets, respectively.

A.2. Segmentation tasks

We utilized a tokenizer chosen from the set $\{\mathcal{T}^{2d}, \mathcal{T}^{3d}\}$ based on the dimensionality of input data, the encoder ϕ, and the randomly initialized decoder and segmentation head to produce prediction maps. Following [1, 5], we devised 2D and 3D versions of decoders and segmentation heads to handle tasks in corresponding dimensions.

Considering a 2D image $I \in \mathbb{R}^{C\times H\times W}$ with resolution (H, W) and C input channels, it is divided into flattened, non-overlapping token sequences $I_{seq} \in \mathbb{R}^{N\times(P^2C)}$. Here, (P, P) represents the resolution of each patch, and $N = (H \times W)/P^2$ denotes the total number of patches. These token sequences are then passed through the encoder, from which we extract four sequence features $\{z_4, z_7, z_{10}, z_{12}\}$, corresponding to the output of the 4-th, 7-th, 10-th, and 12-th layers. These features $z_i \in \mathbb{R}^{\frac{H\times W}{P^2}\times 768}$ are subsequently reshaped to the shape of $768 \times \frac{H}{P} \times \frac{W}{P}$. The bottleneck output of the encoder, *i.e.*, z_{12}, along with the mid-stage output z_{10}, are processed by deconvolutional layers to upscale their resolutions. After concatenation, the obtained features are input into a residual convolution block to produce fused feature maps. Such a similar process is iteratively applied across all subsequent decoder layers up to the original input resolution where the last output is passed through a 1×1 convolutional layer to generate pixel-wise segmentation predictions. More details were displayed in Fig. 1.

For 3D image inputs, we employ a similar architecture, where the 2D tokenizer, convolutional layers, and normalization layers are replaced with their 3D counterparts, allowing for accepting 3D data.

B. Dataset Details

B.1. Details of upstream datasets

JPG version of the MIMIC-CXR 2.0.0 dataset: This extensive public dataset comprises 377,110 JPG format chest radiographs and 227,827 clinical reports associated with these images. Following [6, 11], we excluded all lateral views from the dataset, as the downstream datasets only contain frontal-view chest images. This resulted in a collection of 227,323 clinical reports and 356,309 chest radiograph images.

†Yutong Xie and Yong Xia are the corresponding authors.

https://physionet.org/content/mimic-cxr-jpg/2.0.0/

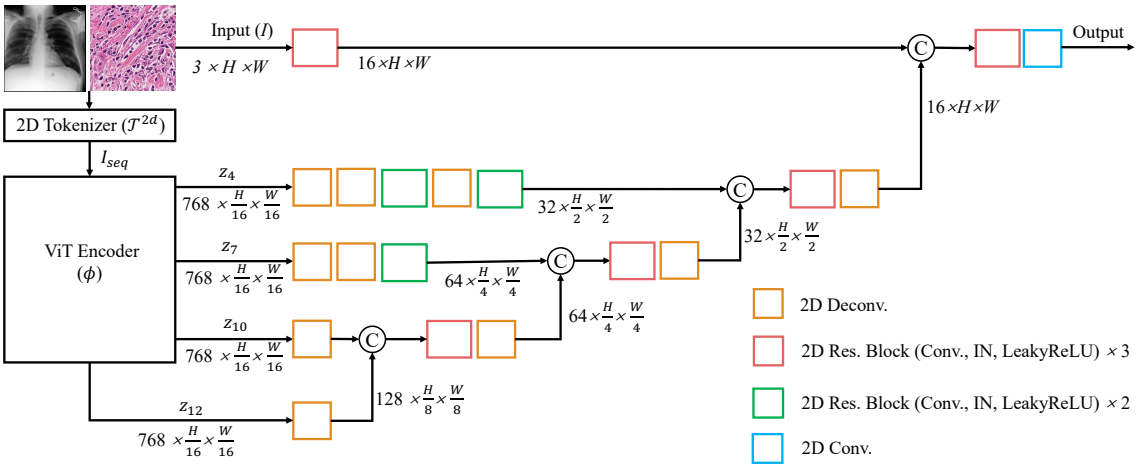

Figure 1. Illustration of model architecture for 2D segmentation tasks.

DeepLesion dataset: This dataset includes 10,594 CT scans collected from 4,427 subjects . Following [8], each CT scan was resampled to a uniform spacing of $1.0 \times 1.0 \times 1.0$. We then applied cropping and resizing strategies to extract 125,070 sub-volumes. Specifically, we cropped each scan along the depth dimension using 24 window lengths and 12 strides, resizing the obtained sub-volumes to $16 \times 192 \times 192$.

ADNI dataset: This dataset is a combination of ADNI-1, ADNI-2, and ADNI-GO datasets, utilizing a defined screening strategy . The selection was based on diagnostic labels such as Alzheimer's disease, mild cognitive impairment, and cognitively normal, without specific consideration for sex, age, slice thickness, and manufacturer. Each MRI scan was cropped along the depth dimension using 16 window lengths and 8 strides, resizing the cropped volumes to $16 \times 192 \times 192$. In total, 59,205 sub-volumes were extracted.

TCGA dataset: Comprising seven projects (TCGA-THYM, TCGA-THCA, TCGA-BRCA, TCGA-UCEC, TCGA-UVM, TCGA-OV, and TCGA-MESO) , this dataset includes a variety of pathological images. We processed these images by cropping them into non-overlapped 512×512 patches and resizing them to 224×224. For each patient with n patches, we randomly selected $min(100, n)$ patches for training data.

B.2. Details of downstream datasets

In Table 1, we provide the implementation details of nine downstream datasets. This includes information on the task type, modality, loss function, patch size, optimizer, learning rate, batch size, and maximum iterations. Additional information on each dataset is as follows: (1) **PudMed20k dataset**. This dataset contains 20,000 abstracts from randomized controlled trials (RCTs), featuring a vocabulary of 68,000 words across 240,000 sentences. Each sentence is categorized into one of five labels: background, objective, method, result, or conclusion. We adopted the official data split, dividing the dataset into training, validation, and test sets, utilizing only the text data for category prediction. (2) **ChestXR dataset**. This dataset focuses on detecting COVID-19, pneumonia, or normal in Chest X-ray images. It includes an official split of 14,958 training images and 3,432 test images, with 20% of the training data randomly selected as the validation set. (3) **QaTa-COV19-v2 (QaTa) dataset**. This dataset is utilized for COVID-19 infected region segmentation, with an official split of 7,145 training and 2,113 test images. A random 20% of the training data serves as the validation set. (4) **RICORD dataset**. This dataset contains 330 CT scans, which are divided into two categories: COVID-19 and normal. All images were resized to $64 \times 192 \times 192$. We followed the data split in [7]. (5) **LiTS dataset**. This dataset includes 131 CT scans with annotations of liver and liver tumor segmentation, following the data split in [9]. The data were preprocessed by using nnUNet's preprocessing procedure [2]. (6) **Vestibular-Schwannoma-SEG (VS) dataset**. This dataset contains 242 MRIs collected on patients with vestibular schwannoma. We followed the data split in [9]. The data were preprocessed by using nnUNet's preprocessing procedure [2]. (7) **LA dataset**. This dataset provides 100 gadolinium-enhanced MRIs paired with left atrium ground truths. We followed the data split and preprocessing steps described in [10]. (8) **NCH dataset**. It consists of the NCT-CRC-HE-100K dataset (training set) and the CRC-VAL-HE-7K dataset (test set). We randomly sample 20% of the training data as the validation set of each category. (9) **GlaS dataset**.

https://nihcc.app.box.com/v/DeepLesion
adni.loni.usc.edu
https://portal.gdc.cancer.gov/

Table 1. Implementation details of nine downstream tasks. CE: cross-entropy loss function.

Dataset	Task Type	Modality	Loss Function	Patch Size	Optimizer	Learning Rate	Batch Size	Iterations
PudMed20k	Cls	1D Report	CE	112	AdamW	0.0002	64	14,065
ChestXR	Cls	2D X-ray	CE	224×224	AdamW	0.00005	32	35,840
QaTa	Seg	2D X-ray	Dice + CE	224×224	AdamW	0.0001	16	25,000
RICORD	Cls	3D CT	CE	$64 \times 192 \times 192$	AdamW	0.00001	8	9,600
LiTS	Seg	3D CT	Dice + CE	$64 \times 192 \times 192$	AdamW	0.0001	2	25,000
VS	Seg	3D MR	Dice + CE	$64 \times 192 \times 192$	AdamW	0.0001	2	25,000
LA	Seg	3D MR	Dice + CE	$64 \times 192 \times 192$	AdamW	0.00005	2	25,000
NCH	Cls	2D Path.	CE	224×224	AdamW	0.0001	32	24,990
GlaS	Seg	2D Path.	Dice + CE	512×512	AdamW	0.0001	4	25,000

Table 2. Fine-tuning performance of five pre-trained models on four datasets, each representing different modalities. DeSD: 3D ResNet pre-trained on CT scans; Path_DINO: ViT pre-trained on pathological images; PCRLv2 (CheXpert/LUNA): 2D and 3D ResNets pre-trained on CheXPert (X-rays) and LUNA (CT scans), respectively; UniMiSS: Dimension-free pyramid U-like medical transformer (MiT) pre-trained on X-rays and CT scans. $\times$ means the model is incompatible with the dataset's data, resulting in an inability to be fine-tuned.

Method	PudMed20k (Report)			ChestXR (X-ray)			RICORD (CT)			NCH (Path.)		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
DeSD [8]	×	×	×	×	×	×	78.57	83.46	83.74	×	×	×
PCRLv2 (LUNA) [12]	×	×	×	×	×	×	81.35	86.24	86.62	×	×	×
Path_DINO [3]	×	×	×	93.00	98.32	92.35	×	×	×	96.10	99.61	94.47
PCRLv2 (CheXpert) [12]	×	×	×	95.41	99.03	94.95	×	×	×	92.99	99.12	89.56
UniMiSS [7]	×	×	×	94.00	98.79	93.46	82.94	87.48	87.52	93.08	99.23	89.99
MedCoSS	83.59	95.38	77.87	94.31	98.83	93.77	83.33	88.74	87.87	95.76	99.51	94.01

Table 3. Performance of different backbones with different pre-trained models. CT: single-modal SSL pre-training on CT data. The best result in each column is highlighted with **bold**.

Method	Backbone	Liver		Liver tumor		Average	
		Dice	HD	Dice	HD	Dice	HD
CT	ViT/B	95.82	7.39	48.13	60.19	71.98	33.79
MedCoSS	ViT/B	95.41	9.54	48.61	63.47	72.01	36.50
DeSD [8]	ResUNet	96.81	4.24	65.27	**26.84**	81.04	15.54
CT + DeSD	ResUNet + ViT/B	96.82	3.67	66.14	26.99	81.48	**15.33**
MedCoSS + DeSD	ResUNet + ViT/B	**96.90**	**3.27**	**66.79**	27.84	**81.84**	15.56

This dataset includes 165 pathological images from H&E-stained colon tissue sections, labeled as malignant or benign. We adhere to the official data split and randomly sample 20% of the training data for each category as the validation set.

C. Comparing to Pre-trained Models

We compared our MedCoSS over recent popular pre-trained models, including DeSD [8], PCRLv2 [12], Path_DINO [3], and UniMiSS [7]. All competing models were employed using their officially released weights and were fine-tuned for specific downstream tasks. The results shown in Table 2 indicate that most models, except for the model from UniMiSS, are limited to handling data of a specific dimension. These models demonstrate a significant decrease in performance when tasked with handling data from a different modality. For instance, Path_DINO demonstrates superior performance across all metrics on the NCH dataset, yet

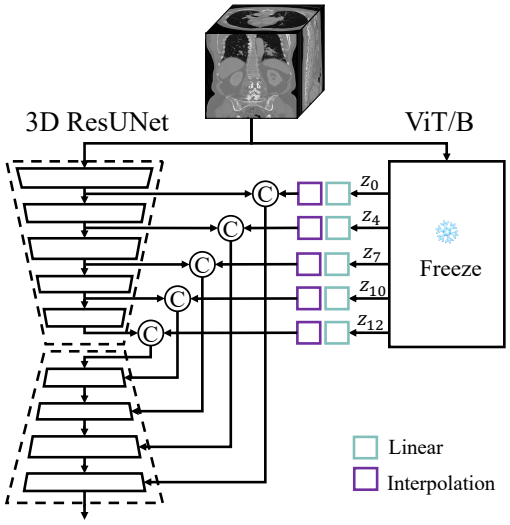

Figure 2. Illustration of the combination of ResUNet and ViT/B. The parameters of ResUNet and ViT/B are learnable and frozen, respectively.

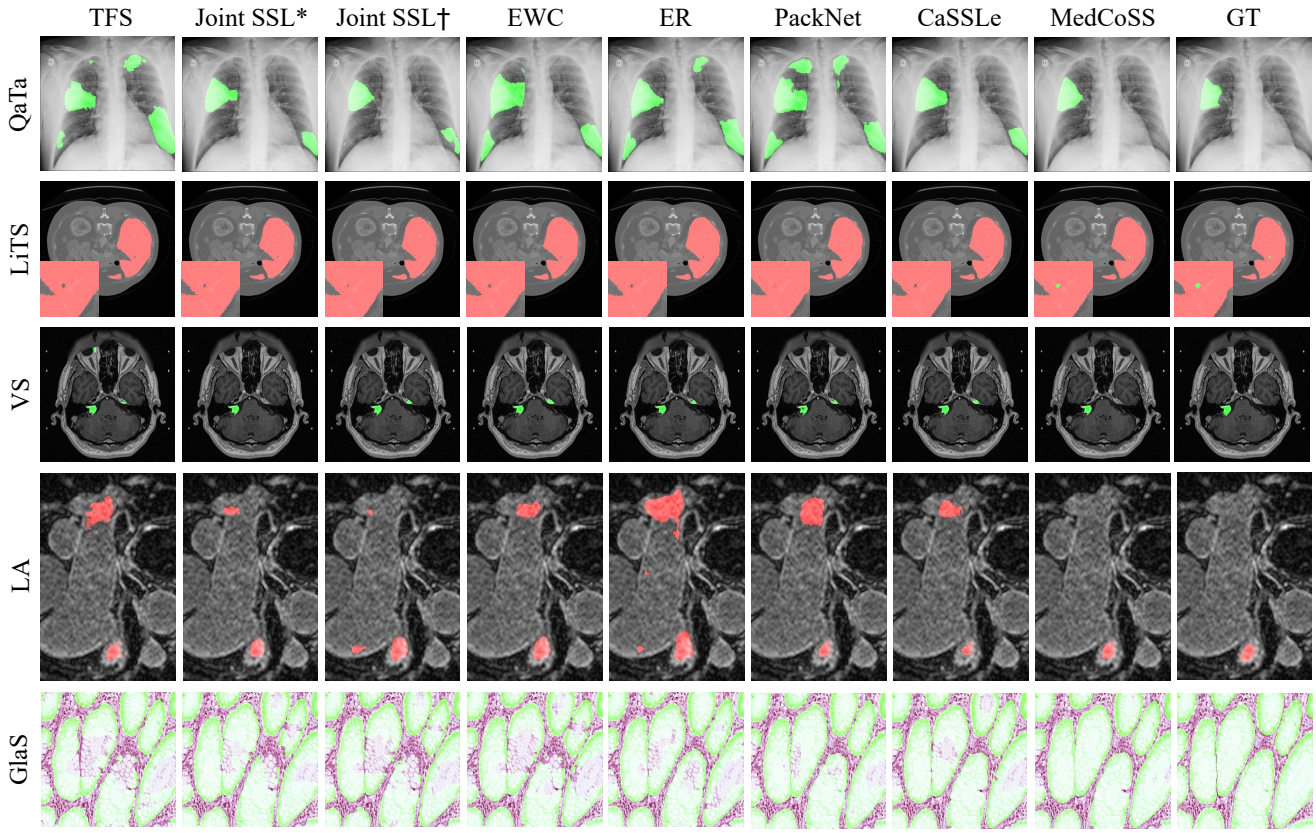

Figure 3. Visualization of segmentation results obtained by TFS, Joint SSL*, Joint SSL†, EWC, ER, PackNet, CaSSLe, and MedCoSS, and ground truths (GTs). The organs are colored red, and the tumors and malignant regions are colored green.

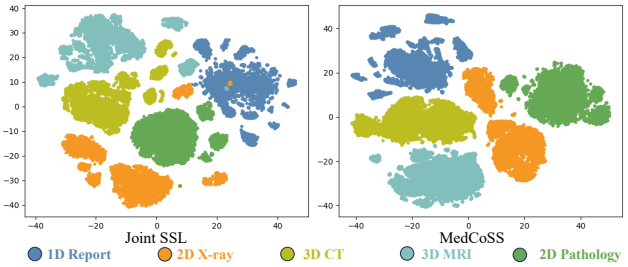

Figure 4. t-SNE visualizations of embeddings from the Joint SSL* and MedCoSS on five modality data.

it obtains the worst performance across all metrics on the ChestXR dataset. In contrast, our MedCoSS is versatile, accepting 1D, 2D, and 3D data, and consistently achieves robust generalization performance across all tasks.

D. Improving 3D Segmentation

While ViT-based models excel in uniformly handling multi-dimensional medical tasks and overcoming dimensional constraints, they tend to underperform in accurately segmenting small targets, such as tumors, which are prone to disappear or lost during passing through the tokenizer module. Inspired by nnSAM [4], we propose an integration of well-pre-trained ViT (obtained from MedCoSS or single-modal SSL) with CNN-based models (obtained from DeSD [8]). This integration strategy, illustrated in Fig. 2, was tested on the LiTS dataset, containing liver and liver tumor segmentation annotations. The results in Table 3 reveal that the pre-trained ResUNet obtained from DeSD achieves a significant performance gain compared to ViT-based models pre-trained by MedCoSS or single-modal SSL (CT). More importantly, the combination of ViT and ResUNet leads to further enhancement in performance. This finding highlights the effectiveness of the integration, suggesting an alternative approach for leveraging the advantages of multi-modal pre-training to enhance the model's segmentation performance.

E. Visualization of Segmentation Results

For a qualitative comparison, we visualized the segmentation results derived from TFS, Joint SSL*, Joint SSL†, EWC, ER, PackNet, CaSSLe, and MedCoSS across five datasets, as shown in Fig. 3. These visualizations reveal that

the segmentation results from MedCoSS most closely align with the ground truths (GTs), effectively avoiding issues of over-segmentation and under-segmentation. For example, as shown in the second row of Fig. 3, MedCoSS uniquely identifies a small liver tumor in the CT images, a task where the other paradigms fail.

F. Visualization of *Modal Data Collision*

We visualized the distribution of five pre-training modalities obtained by Joint SSL (with dimension-shared decoders) and MedCoSS in Fig. 4, where 10K samples were randomly selected from each modality. It shows that, for Joint SSL, the embeddings of different modalities are dispersed and lack clear classification boundaries, indicating that Joint SSL is limited in distinguishing and capturing the unique features of each modality. In contrast, MedCoSS's visualization displays a distinct and compact cluster for each modality, with clear boundaries. It suggests that MedCoSS is more adept at recognizing and preserving the distinct characteristics of each modality. The visualization of Joint SSL intuitively verifies the concept of *modal data collision*. This collision can hinder the model's ability to distinguish each modality, which is a challenge that MedCoSS addresses effectively.

References

[1] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022. 1

[2] 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. 2

[3] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *CVPR*, pages 3344–3354, 2023. 3

[4] Yunxiang Li, Bowen Jing, Xiang Feng, Zihan Li, Yongbo He, Jing Wang, and You Zhang. nnsam: Plug-and-play segment anything model improves nnunet performance. *arXiv preprint arXiv:2309.16967*, 2023. 4

[5] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *ICCV*, pages 21152–21164, 2023. 1

[6] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*, 35:33536–33549, 2022. 1

[7] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575. Springer, 2022. 2, 3

[8] Yiwen Ye, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. In *MICCAI*, pages 545–555. Springer, 2022. 2, 3, 4

[9] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. *arXiv preprint arXiv:2304.03493*, 2023. 2

[10] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, pages 605–613. Springer, 2019. 2

[11] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Mach. Learn. Healthc. Conf.*, pages 2–25. PMLR, 2022. 1

[12] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *TPAMI*, 2023. 3