UniMiSS: Universal Medical Self-Supervised Learning via Breaking Dimensionality Barrier

Yutong Xie[1[0000−0002−6644−1250]], Jianpeng Zhang[2], Yong Xia[2], and Qi Wu[1⋆]

[1] The University of Adelaide, Australia
[2] School of Computer Science and Engineering, Northwestern Polytechnical University, China
yutong.xie678@gmail.com;qi.wu01@adelaide.edu.au

Abstract. Self-supervised learning (SSL) opens up huge opportunities for medical image analysis that is well known for its lack of annotations. However, aggregating massive (unlabeled) 3D medical images like computerized tomography (CT) remains challenging due to its high imaging cost and privacy restrictions. In this paper, we advocate bringing a wealth of 2D images like chest X-rays as compensation for the lack of 3D data, aiming to build a universal medical self-supervised representation learning framework, called UniMiSS. The following problem is how to break the dimensionality barrier, *i.e.*, making it possible to perform SSL with both 2D and 3D images? To achieve this, we design a pyramid U-like medical Transformer (MiT). It is composed of the switchable patch embedding (SPE) module and Transformers. The SPE module adaptively switches to either 2D or 3D patch embedding, depending on the input dimension. The embedded patches are converted into a sequence regardless of their original dimensions. The Transformers model the long-term dependencies in a sequence-to-sequence manner, thus enabling UniMiSS to learn representations from both 2D and 3D images. With the MiT as the backbone, we perform the UniMiSS in a self-distillation manner. We conduct expensive experiments on six 3D/2D medical image analysis tasks, including segmentation and classification. The results show that the proposed UniMiSS achieves promising performance on various downstream tasks, outperforming the ImageNet pre-training and other advanced SSL counterparts substantially. Code is available at https://github.com/YtongXie/UniMiSS-code.

Keywords: Self-supervised Learning; Cross-dimension; Medical Image Analysis; Transformer

1 Introduction

Medical image analysis, a key process in computer-aided diagnosis, is well known by its lack of labels for training, especially for the 3D task. Recent research work suggests that the self-supervised learning (SSL) is promising to ease the

⋆ Corresponding author.

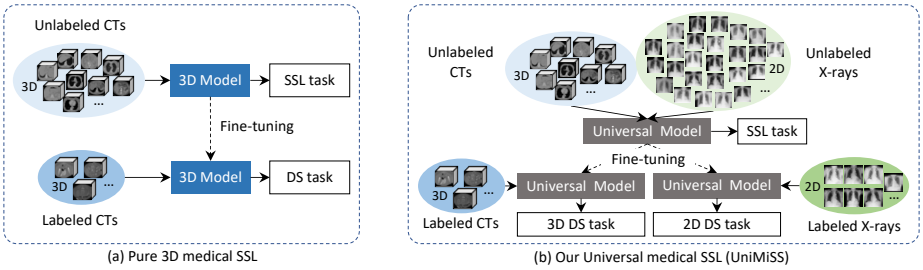

Fig. 1. (a) Pure 3D medical SSL learns representations with only 3D CT scans; (b) our proposed UniMiSS brings a wealth of 2D X-rays to offset the lack of 3D data, thus enables the large-scale SSL for better pre-training performance. Besides, the pre-trained model is generic to various downstream (DS) applications, without the restriction on the dimensionality barrier.

annotation cost by making the best of unlabeled data [8,9,40,41,47,55,57,58]. Although setting label free, SSL still heavily relies on the large-scale unlabeled data to explore the feature representations. Unfortunately, publicly available 3D medical data is relatively limited due to the high imaging cost and data privacy. Most of 3D medical datasets just contain a few thousands of cases. For example, Zhou *et al*. [55] utilized the LUNA dataset [38], containing about 1000 CT cases, for self-supervised pre-training. Such a small data scale may limit the potential of SSL in 3D medical image analysis.

In comparison to 3D data, it is easy to collect hundreds of thousands of 2D medical images such as X-rays due to its fast imaging speed, low radiation and low cost. Accordingly, we advocate to bring a wealth of 2D medical images to the 3D SSL process, aiming at learning strong representations with large-scale images, as shown in Fig. 1. Comparing to the pure 3D medical SSL, this practice benefits the medical SSL in terms of three significant merits. First, 2D data serves as a compensation for the lack of 3D data, enabling the large-scale SSL pre-training. Second, there is the anatomy correlation between 2D and 3D images, like chest X-ray and CT. Such an intrinsic relevance may contribute for strong associated representations. Third, the pre-trained model is generic enough to be applied to both 3D and 2D downstream tasks. To achieve the universal SSL purpose, on the technical side, we need to build a versatile model that is able to process both 2D and 3D images. The common practice in medical image analysis is to design 2D convolutional neural networks (CNNs) for 2D images [49,51,55] and 3D CNNs for 3D images [47,48,52,55,57], respectively. Restricted to the dimensionality barrier, it is almost impossible to design a dimension-free CNN network for this purpose.

Recent months have witnessed the success of Transformer in computer vision [15]. A vision Transformer usually takes a sequence of image patches, represented by the learned linear embedding, as the input to model the long-term dependencies among the sequence elements. Owing to the sequence modeling, Transformer can accept the data of any dimensions, including but not limited to

2D images and 3D volumetric data. Therefore, Transformer offers the possibility of breaking the dimensionality barrier and constructing a universal SSL model.

In this paper, we propose a **Uni**versal **M**ed**i**cal **S**elf-**S**upervised representation learning framework (UniMiSS) that learns general representations from 2D and 3D unlabeled medical images. To achieve this, we design a dimension-free pyramid U-like **M**ed**i**cal **T**ransformer (MiT), which is mainly composed of switchable patch embedding (SPE) module and Transformers. The SPE module converts the input images to a sequence by using 2D or 3D patch embedding, depending on the input dimension. The Transformer layer processes the embedded tokens in a sequence-to-sequence manner, regardless of their original dimension. We perform the self-supervised learning by the self-distillation of student and teacher networks, both of which take the MiT as the backbone. The student network learns to predict the output distribution obtained with the momentum teacher network, following the view consistency. Moreover, the 3D volumetric image should be identical with their slices due to the same imaging content. The volume-slice consistency is adopted as a cross-dimension regularization to boost the representations. We conduct the SSL experiments based on 5,022 3D CT volumes, which are augmented by 108,948 2D X-ray images. Benefit from the huge augmented 2D data, the proposed UniMiSS achieves the obvious performance improvement on the downstream 3D classification/segmentation tasks. Besides, the UniMiSS pre-trained model can be freely applied to 2D downstream tasks, which beats strong competitors like ImageNet pre-training on the downstream 2D medical tasks.

To summarise, our contributions are three-fold: (1) we are the first to augment 3D medical images with the easily accessible unpaired 2D ones for the SSL purpose, aiming at addressing the limitation of 3D data amounts during the SSL process; (2) the proposed MiT breaks the dimensionality barrier and enables the joint SSL training with both 2D and 3D images; and (3) our UniMiSS pre-training achieves the advanced performance on six downstream tasks, covering the 3D/2D medical image classification/segmentation.

2 Related Work

2.1 Self-supervised Learning

SSL has been extensively studied in the literature. According to the pretext tasks, these studies can be broadly categorized into the discriminative methods [6,11,17,19,21,28,33,34,35,42] and generative methods [26,27,36,37,53]. The contrastive learning [11,17,19,21,33,35,42] has drawn significant research attention and achieved advanced performance on many vision tasks. Most of the previous work were built on the CNN-based network. More recently, Transformer has become an increasingly popular alternative architecture in computer vision. There has been a trend towards combining the merits of Transformer and SSL, advancing the self-supervised vision Transformers. The seminal work is iGPT [10], which follows the masked auto-regressive language modelling to pre-train the self-supervised vision Transformer. Besides, some attempts have also

been made to pre-train vision Transformers using the contrastive learning [7] or Siamese distillation [13], which outperform the CNN-based SSL approaches, setting a new record on ImageNet.

The success of SSL in computer vision also benefits to the medical community [8,9,40,41,47,55,57,58]. Typical attempts include pre-training a CNN by restoring the content of raw images [9,41,55,57,58] and tailoring contrastive SSL to medical images [8,40,41,47]. These efforts constitute an important and timely step forward towards better SSL approaches to medical image analysis. However, they suffer two limitations. First, the CNN architecture enables the pre-training on either 2D or 3D medical images, failing to process both of them simultaneously. The resulting representations would be trapped especially for the limited 3D data. Consequently, the pre-trained CNN can only be transferred to the dimension-specific downstream task. Second, the above SSL approaches capture the spatial context of 3D medical images from either slices [8] or volume [41,57,58]. Few of them consider the inherent consistency relation between volume and its slices.

2.2 Cross-Domain Training for Medical Imaging

In the medical context, the cross-domain training usually jointly utilizes two or more datasets acquired at different sites [24,30] or using different imaging modalities [16,29,54] to train a single model that could perform well on diverse datasets. Karani *et al.* [24] and Liu *et al.* [30] trained a single CNN with shared convolutional layers and specific batch normalization layers using the MRI data acquired at each site individually, aiming to tackle the statistical divergence explicitly. Zhang *et al.* [54] simultaneously learned a volume-to-volume translation using the unpaired CT and MRI data and strong segmentors using synthetic data, which were translated from another modality. Dou *et al.* [16] derived a variant of knowledge distillation (KD) to leverage the shared across-modality information between CT and MRI for accurate segmentation of anatomical structures. Li *et al.* [29] also introduced KD to the cross-modality analysis of CT and MRI data, but they simultaneously exploited abundant unlabeled data. These studies are dedicated to analyzing multi-modal/site but fixed dimension (3D) medical images, failing to address the dimensionality barrier in our scenario.

3 Methods

3.1 Overview

UniMiSS is a universal medical SSL framework that is superior to learn general image representations with large scale mixed 2D and 3D unlabeled medical images. Figure 2 illustrates the pipeline of UniMiSS. Let us denote the mixed 2D and 3D data pool by $\{\mathbb{D}^{2D}, \mathbb{D}^{3D}\}$. To enable UniMiSS to process both 2D and 3D medical images, we build the MiT as its backbone, which is mainly constituted by the dimension-adaptive SPE module and Transformer layers. We

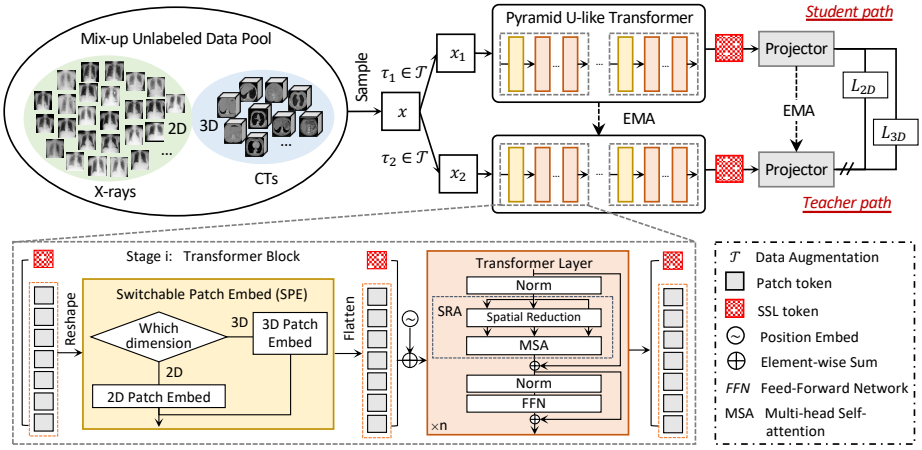

Fig. 2. Illustration of the proposed UniMiSS framework. It has a dual path architecture, *i.e.*, a student and a teacher. Taking both 2D X-rays and 3D CTs as input, UniMiSS is trained by the self-distillation strategy, *i.e.*, maximizing the agreement of both paths. To break the dimensionality barrier between X-rays and CTs, the MiT network, composed of the switchable patch embedding (SPE) module and Transformers, processes the 3D/2D data in a sequence-to-sequence manner.

perform the SSL process in the self-distillation manner, and utilize a standard cross-entropy loss to maximize the consistency between the student and teacher outputs. Besides, to get the utmost out of 3D volumetric information, we introduce the volume-slice consistency constraint, which encourages UniMiSS to model the consistency cross dimensions. It is intuitively conducive to learning strong feature representations from the volumetric images. We now delve into the details of this framework.

3.2 MiT: A Dimension-free Architecture

Although achieving great success in computer vision, vision Transformer [15] still remains challenging to process high resolution 3D images, due to the high computation cost and memory requirement. Inspired by [45], we design the MiT with a pyramid architecture to process both 2D and 3D images efficiently. To break the dimensionality barrier, we propose a simple yet efficient SPE module to adaptively choose the 2D or 3D patch embedding according to the input type. MiT has an encoder-decoder architecture that facilitates the various applications, including segmentation and classification. We now describe each part of MiT, and more details can be found in Appendix.

SPE. As shown in Figure 2, the SPE module plays an important role to obtain the dimension-specific embedding, *i.e.*, using 2D patch embedding operation for 2D inputs and using 3D patch embedding operation for 3D inputs . Notice that the implementations of SPE in the encoder and decoder are different. The SPE in the encoder refers to a switchable 2D and 3D convolution block with the stride

of 2, which reduces the feature resolution. In contrast, the SPE in the decoder is a switchable 2D and 3D transpose convolution block, which increases the feature resolution.

Encoder-Decoder. The MiT encoder follows a progressive shrinking pyramid Transformer, as done in [45]. It consists explicitly of four stages, each of which is composed of a SPE module and several stacked Transformers. In each stage, the SPE module down-samples the input features and generates the dimension-specific embedded sequence. Notably, we append an extra learnable SSL token [7,13] to the patch embedded sequence. The SSL token is similar to the [CLS] token in ViT, which is able to aggregate information from the whole patch embedding tokens via the self-attention. The resultant sequences, combined with the learnable positional embedding, are inputted into the following Transformers for the long-term dependency modeling. Each Transformer layer includes a self-attention module and a feed-forward network (FFN) with two hidden layers. To enable MiT to process high-resolution images, we follow the spatial-reduction attention (SRA) layer [45]. Given a query $\boldsymbol{q}$, a key $\boldsymbol{k}$, and a value $\boldsymbol{v}$ as the input, SRA first reduces the spatial resolution of $\boldsymbol{k}$ and $\boldsymbol{v}$, and then feeds $\boldsymbol{q}$, reduced $\boldsymbol{k}$, and reduced $\boldsymbol{v}$ to a multi-head self-attention (MSA) layer to produce refined features. This process can be formally expressed as follows

$$SRA(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = MSA(\boldsymbol{q}, F(\sigma(R(\boldsymbol{k}))), F(\sigma(R(\boldsymbol{v})))), \tag{1}$$

where $\sigma(\cdot)$ represents a linear projection, *i.e.*, strided 2D or 3D convolution operation, that reduces the feature map resolution, $R(\cdot)$ reshapes the input sequence to a feature map of the original spatial size, and $F(\cdot)$ flattens the input into a 1D sequence. MiT has a symmetric decoder structure that consists of three stages. In each stage, the input feature map is first up-sampled by the SPE module, and then refined by the stacked Transformer layers. Besides, we also add skip connections between the encoder and decoder to keep more low-level but high-resolution information.

3.3 Objective of UniMiSS

The proposed UniMiSS framework is based on the student-teacher paradigm. Each path comprises a MiT network $\mathcal{F}_\theta(\cdot)$ and a projector $\mathcal{P}_\theta(\cdot)$. $\mathcal{P}_\theta(\cdot)$ is a n-layer multi-layer perceptron (MLP) head, θ represents the parameter set of this path. The SPE layers switch to perform the 2D patch embedding or 3D patch embedding during the feed-forward computing that is denoted as $\mathcal{F}_\theta(\cdot; 2D)$ and $\mathcal{F}_\theta(\cdot; 3D)$, respectively. During the SSL process, we only extract the SSL token from the output of $\mathcal{F}_\theta(\cdot; 2D/3D)$ as the input of the projector. Since the Transformer sets the dimension free, our UniMiSS is able to learn image representations from both 2D and 3D unlabeled medical images.

Both of paths share an identical architecture. However, they differ in the following two items. First, the teacher network is formulated as a momentum version of the student network, which updated by an exponential moving average strategy, defined as

$$\mu \leftarrow \lambda\mu + (1 - \lambda)\theta, \tag{2}$$

where λ increases from 0.996 to 1 using a cosine schedule during training [7]. Second, a stop-gradient operator is performed to the teacher network to avoid model collapse.

Objective for 2D domain data. Taking a mini-batch of 2D data $\boldsymbol{x}$ for example, we first create two augmented views $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ by using the data augmentation module $\mathcal{T}$, and then feed them into the student and teacher networks. The obtained SSL token is inputted into the projector to produce the output vector, denoted as $\boldsymbol{f}_1 = \mathcal{P}_\theta(\mathcal{F}_\theta(\boldsymbol{x}_1; 2D))$, $\boldsymbol{f}_2 = \mathcal{P}_\mu(\mathcal{F}_\mu(\boldsymbol{x}_2; 2D))$. The objective of UniMiSS is to maximize the consistency between the output vectors obtained with student and teacher networks, formulated by

$$\mathcal{H}(\boldsymbol{f}_1, \boldsymbol{f}_2) = -\text{softmax}(\frac{\boldsymbol{f}_2 - \mathcal{C}}{\tau_t}) * \log(\text{softmax}(\frac{\boldsymbol{f}_1}{\tau_s})), \tag{3}$$

where $\mathcal{C}$ is the centering of teacher outputs, τ_t and τ_s are sharpening temperature parameters for student and teacher network. The centering operation heartens the model to the uniform distribution while the sharpening has the opposite effect, *i.e.*, encouraging one dimension to dominate. Both of them are jointly used together to avoid model collapse [7]. Specifically, the temperature τ_t is set to a small value in the teacher path for the sharpening purpose. The center $\mathcal{C}$ is first computed via averaging the teacher's outputs of the min-batch data and then updated with an exponential moving average strategy to aggregate the center across the whole batches, shown as follows

$$\mathcal{C} \leftarrow \omega * \mathcal{C} + (1 - \omega) * \widehat{\boldsymbol{f}_2} \tag{4}$$

where ω is a rate parameter, and $\widehat{\boldsymbol{f}_2}$ refers to the mean of teacher output in a mini-batch. We define a symmetrized loss for 2D images as:

$$\mathcal{L}^{\text{2D}} = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{D}^{2D}}[\mathcal{H}(\boldsymbol{f}_1, \boldsymbol{f}_2) + \mathcal{H}(\boldsymbol{f}_2, \boldsymbol{f}_1)] \tag{5}$$

Objective for 3D domain data. In medical domain, 3D volumes can be viewed as the stacking of 2D images along with the inter-slice dimension. The volume data has the inherent consistency to their slices, which inspires us to model the volume-slice consistency for SSL. Given a 3D data $\boldsymbol{x}$ sampled from the 3D medical dataset, we denote its two augmented views as $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, each containing m 2D slices. We compute the global volumetric representations by the student and teacher networks in a 3D mode, *i.e.*, $\boldsymbol{f}_1 = \mathcal{P}_\theta(\mathcal{F}_\theta(\boldsymbol{x}_1; 3D))$, and $\boldsymbol{f}_2 = \mathcal{P}_\mu(\mathcal{F}_\mu(\boldsymbol{x}_2; 3D))$. Meanwhile, we stack m slices of each augmented view in a batch, and use them as 2D inputs to calculate the slice-wise representations in a 2D mode, and then treat the average outputs of all slices as the holistic slice representations, *i.e.*, $\boldsymbol{f}_1' = \frac{1}{m}\sum_{i=1}^{m} \mathcal{P}_\theta(\mathcal{F}_\theta([\boldsymbol{x}_1]^i; 2D))$, and $\boldsymbol{f}_2' = \frac{1}{m}\sum_{i=1}^{m} \mathcal{P}_\mu(\mathcal{F}_\mu([\boldsymbol{x}_2]^i; 2D))$, where $[\boldsymbol{x}]^i$ represents the i-th slice extracted from the 3D data $\boldsymbol{x}$. After that, we build the following objective function

$$\begin{aligned}\mathcal{L}^{\text{3D}} = \mathbb{E}_{\boldsymbol{x} \sim \mathbb{D}^{3D}}[&\mathcal{H}(\boldsymbol{f}_1, \boldsymbol{f}_2) + \mathcal{H}(\boldsymbol{f}_1, \boldsymbol{f}_2') + \mathcal{H}(\boldsymbol{f}_1', \boldsymbol{f}_2) + \mathcal{H}(\boldsymbol{f}_1', \boldsymbol{f}_2') \\ &+ \mathcal{H}(\boldsymbol{f}_2, \boldsymbol{f}_1) + \mathcal{H}(\boldsymbol{f}_2, \boldsymbol{f}_1') + \mathcal{H}(\boldsymbol{f}_2', \boldsymbol{f}_1) + \mathcal{H}(\boldsymbol{f}_2', \boldsymbol{f}_1')]\end{aligned} \tag{6}$$

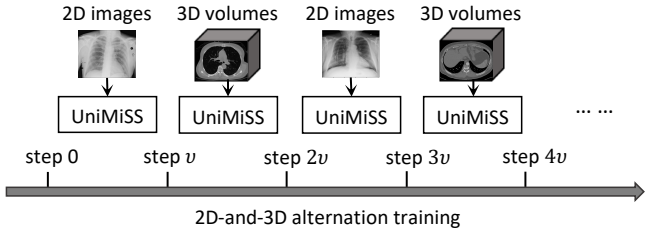

Fig. 3. Illustration of 2D-and-3D alternation training.

The above objective function encourages to learn the refined consistency with 3D medical data in terms of three aspects, *i.e.*, volume to volume, slice to slice, and volume to slice.

We introduce an alternative training scheme to solve this multi-objective optimization problem. As shown in Figure 3, we first sample 2D images to train the UniMiSS from step 0 to step υ, and then take turn to sample 3D volumes in the next υ steps. The following training process will continue in a circular manner until the model converges. The proposed iterative training scheme has two merits: (1) it bypasses the difficulty of using both 2D and 3D images in the same batch; and (2) it can reduce the instability caused by the distribution discrepancy between 2D and 3D data.

4 Experiments

4.1 Datasets

Pre-training datasets. We collected 5,022 3D CT scans from five datasets (*i.e.* MOTS dataset [52], LIDC-IDRI dataset [5], Tianchi dataset [2], RibFrac dataset [23], TCIACT dataset [4]), and collected 108,948 2D images from NIH ChestX-ray8 dataset [46] to train UniMiSS in a self-supervised manner.

Downstream datasets. Table 1 gives the details of six downstream tasks, which can be grouped into (1) 3D downstream: CT-based segmentation (BCV) and classification (RICORD), MRI-based segmentation (CHAOS); (2) 2D downstream: multi-organ segmentation (JSRT) and pneumonia classification (ChestXR), and skin lesion segmentation (ISIC). Note that the CHAOS and ISIC datasets are different from the pre-training data in terms of modalities (*i.e.*, 3D CT vs. MRI, 2D X-ray vs. dermoscopy). They are used to evaluate the unseen-modality transferability.

4.2 Experimental Details

Pre-training setup. We set the size of input 2D patches to 224×224 and 3D patches to $16 \times 96 \times 96$, aiming to weigh the balance between reserving enough information for SSL and reducing computational and spatial complexity to an affordable level. We applied a rich set of data augmentations to create positive

Table 1. Six datasets for the downstream evaluation. Noticed that we used two test sets, *i.e.* offline test set (off) and online test set (on), for the BCV dataset.

Downstream evaluation datasets				
Name	Tasks	Modalities	#Train	#Test
BCV [1]	Multi-organ segmentation	3D CT	24	6 (off)+20 (on)
RICORD [43]	COVID-19 screening		182	45
CHAOS [25]	Abdominal organ segmentation	3D MRI	48	12
JSRT [39,44]	Multi-organ segmentation	2D X-ray	124	123
ChestXR [3]	Pneumonia classification		17,955	3,430
ISIC [14]	Skin lesion segmentation	2D dermoscopy	2000	600

views, including colour jittering, Gaussian blur/noise, random crop, zooming, and flip to the inputs for producing two views. Following [7], we adopted the AdamW optimizer [32] with a cosine decaying learning rate [31], a warm-up period of 10 epochs, to train our UniMiSS. We empirically set the initial learning rate to 0.0008, batch size to 192, maximum epochs to 200, rate parameter ω to 0.9, and temperature parameter τ_t and τ_s to 0.04 and 0.1, respectively. It took about 2.5 days to pre-train the UniMiSS using 8 NVIDIA V100 GPUs. We understand this is a big GPU consumption but it saves large amount of time and money to collect 3D medical image data, as we use easily-collected 2D data as the fuel.

Downstream training setup. For the classification, we extracted the pre-trained MiT encoder and appended a FC layer with the output channel as the number of classes for prediction. For the segmentation, we took the pre-trained MiT encoder and decoder while removing the SSL token, and appended a segmentation head for prediction. This head includes a transposed convolutional layer, a Conv-IN-LeakyReLU, and a convolutional layer with the kernel size of 1 and the output channel as the number of classes. The segmentation performance is measured by the Dice coefficient scores. The classification performance is measured by the area under the receiver operator curve (AUC). Note that we randomly split 25% training samples as a validation set to select the hyper-parameters of UniMiSS in the ablation study. The detailed training setups for each downstream task are shown in Appendix.

4.3 Results on 3D downstream tasks

Dimension-specific SSL *vs.* Cross-dimension SSL. In this section, we evaluate the SSL performance on two downstream 3D taks, *i.e.*, multi-organ segmentation (BCV) and COVID-19 screening (RICORD). The UniMiSS pre-training is compared with the random initialization (Rand. init.) and five advanced SSL methods, including MoCo v2/v3 [12,13], PGL [47], PCRL [55], and DINO [7]. Note that MoCo v2, PGL, and PCRL take the CNN as their encoder backbone, *i.e.*, a 3D ResNet with 50 learnable layers. During the SSL process, MoCo v2 and PGL only pre-train the encoder part, while PCRL additionally pre-trains

Table 2. Segmentation and classification performance of using different pre-training strategies on the BCV offline test set and RICORD test set.

Methods	Backbone	BCV (CT, seg)			RICORD (CT, cls)		
		20%	40%	100%	20%	40%	100%
Rand. init.	CNN	68.44	73.14	79.93	69.72	74.66	83.36
MoCo v2 [12]		71.22	75.09	82.05	73.46	77.81	85.46
PGL [47]		72.05	75.86	82.57	73.76	77.96	85.61
PCRL [55]		72.80	76.05	82.73	75.11	79.01	86.21
Rand. init.	Transformer	70.09	74.60	79.97	71.36	76.06	83.21
MoCo v3 [13]		74.54	78.16	82.02	75.56	79.66	85.16
DINO [7]		75.33	78.88	82.61	76.31	80.11	85.91
UniMiSS (Ours)		**77.96**	**80.97**	**84.99**	**78.71**	**82.96**	**89.06**

a decoder by using the reconstruction task. Besides, MoCo v3, DINO, and our UniMiSS use the Transformer model as the backbone, which contains both encoder and decoder. We employ the U-like PVT as the backbone for MoCo v3 and DINO, which has a similar architecture of MiT but the different patch embedding module. The lack of SPE make them fail to process both 2D and 3D images simultaneously, resulting in the dimension-specific SSL with only 3D data. For a fair comparison, all of these SSL methods are pre-trained on the 5,022 unlabeled 3D CT scans. Somewhat differently, the proposed UniMiSS introduces the additional 2D X-rays to the 3D SSL training, benefiting from the universality. We make more detailed comparisons between the proposed UniMiSS and other dimension-specific CNN/Transformer SSL methods. Table 2 shows the results of three label settings (20%,40%, and 100% label available). We summarize this table in the following points: (1) The Transformer-based models outperform obviously the CNN-based methods, mainly owing to the SSL pre-training. It reflects that the Transformer is a competitive architecture and the SSL pre-training is essential for the Transformer to achieve good performance. (2) The proposed UniMiSS is superior to MoCo v3 and DINO. The performance gains over DINO are +2.38% for segmentation and +3.15% for classification when 100% labels are available. It proves the effectiveness of using a wealth of 2D medical images to assist the 3D SSL process. (3) Besides, it is really encouraging to see that the proposed UniMiSS is able to achieve the comparable or even superior performance while less annotations, even a half. Taking BCV for example, UniMiSS with 40% label achieves 80.97% segmentation Dice, which is better than the 79.97% of the random initialized method with 100% labels.

Comparisons on the BCV online test set. To be more persuasive, we also compared the proposed UniMiSS with other state-of-the-art segmentation methods on the BCV online test set. As listed in Table 3, these compared methods include PaNN [56], UNETR [18], nnUnet [22] and DoDnet [52]. Note that the performance records of these competitors come from their original paper. It reveals that our UniMiSS, without using any ensemble strategy, still achieves the competitive performance, the best Hausdorff distance (HD) and average mean

Table 3. Comparisons on the BCV online test set.

Metrics	PaNN [56]	UNETR [18]	nnUnet [22]	DoDnet [52]	UniMiSS	UniMiSS
Ensemble	5	5	10	5	1	10
Dice	85.00	85.55	87.62	86.44	87.05	**88.11**
HD	18.47	\	\	15.62	13.92	**13.17**
SD	1.45	\	\	1.17	1.02	**0.90**

Table 4. Segmentation and classification performance of using different pre-training strategies on two 2D test sets.

Methods	Backbone	JSRT (X-ray, seg)			ChestXR (X-ray, cls)		
		20%	40%	100%	20%	40%	100%
Rand. init.	CNN	84.05	87.63	90.96	92.05	94.83	97.54
INpre [20]		87.90	90.01	91.73	94.78	96.26	98.13
MoCo v2 [12]		88.65	91.03	92.32	95.22	96.61	98.67
PGL [47]		89.01	91.39	92.76	95.56	96.96	98.87
PCRL [55]		89.55	91.53	93.07	95.88	97.43	98.99
Rand. init.	Transformer	85.55	88.83	91.22	92.80	95.20	97.04
MoCo v3 [13]		90.07	91.75	92.68	95.99	97.33	98.59
DINO [7]		90.40	92.16	93.03	96.44	97.69	98.70
UniMiSS		**91.88**	**93.15**	**94.08**	**97.09**	**98.14**	**99.07**

surface distance (SD), and second highest Dice on the online test set, outperforming the DoDNet with supervised pre-training. When using the coarse-to-fine ensemble strategy like [22], our UniMiSS can obtain the best performance in terms of all metrics.

Results on 2D downstream tasks. Since pre-training on both 2D and 3D medical images, our UniMiSS can be freely applied to 2D downstream tasks. Table 4 makes the comparisons on the 2D medical image segmentation and classification tasks. The compared methods include the Rand. init., ImageNet pre-training (INpre), CNN-based SSL methods (*i.e.* MoCo v2, PGL and PCRL), and Transformer-based SSL methods (*i.e.* MoCo v3 and DINO). Different from the 3D scenarios, a 2D ResNet-50 is used as the backbone in MoCo v2, PGL, and PCRL. MoCo v3, and DINO still take the U-like PVT as the backbone, but modify the patch embedding to adapt for the 2D inputs. Here, all compared SSL methods are pre-trained on the same 2D unlabeled medical images. As for UniMiSS, we directly apply the previous pre-trained model to the 2D tasks, without any modification or further re-training. From the results, we can find that (1) the SSL methods have surpassed INpre in both tasks, revealing that pre-training on a large-scale medical image dataset is more friendly to medical domain downstream tasks than pre-training on natural images; (2) although the number of 3D data is much smaller than 2D, *i.e.*, about one in twenty, the UniMiSS pre-training still achieves the performance gain over the pure 2D SSL method, like DINO. This may account in part for the inherent correlation between X-rays and CTs. Such a correlation information can be captured by the UniMiSS, thus contributed for the performance gain.

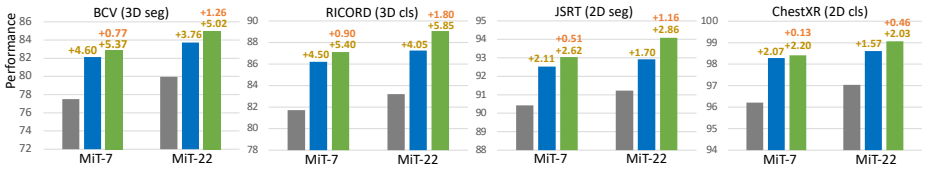

Fig. 4. Results of MiT with fewer Transformer layers. Here, MiT-7 and MiT-22 denote MiT with 7 and 22 Transformer layers, respectively. Rand. init., Dimension-specific pre-training, UniMiSS. Note that the performance gain with yellow and orange color is computed by comparing to the Rand. init., and dimension-specific pre-training baseline, respectively.

Table 5. Segmentation and classification performance on two 3D validation sets with or without using volume-slice consistency.

Objective for 3D		BCV (seg)			RICORD (cls)		
Volume	Slices	20%	40%	100%	20%	40%	100%
✓		72.08	76.04	80.94	69.87	74.61	80.96
✓	✓	**74.56**	**77.97**	**82.36**	**72.46**	**76.89**	**82.43**

4.4 Discussions

Effectiveness of volume-slice consistency. We design the volume-slice consistency mechanism for learning rich representations with 3D medical images. To evaluate the effectiveness of this mechanism, we pre-trained UniMiSS on 3D medical images with or without using the volume-slice consistency. Table 5 gives the downstream performance on the validation of two 3D datasets. The proposed volume-slice consistency can substantially improve the 3D segmentation/classification accuracy under different label ratios. The performance gain is at least by 1.42% on segmentation and by 1.47% on classification.

Number of iteration interval. The UniMiSS is optimized in a 2D-3D alternation training way, where the iteration interval υ is a critical parameter. A smaller υ may lead to insufficient training for each domain. A larger υ may make the network forget the information learned from another domain. To set a suitable υ, we pre-trained UniMiSS with various of υ, varying from 1 to 3, and fine-tuned them on four downstream tasks. Table 6 shows that the pre-trained UniMiSS can achieve the best performance on four downstream tasks when υ equals 2, and below or above 2 gives rise to the performance loss. Hence, we suggest setting the iteration interval to 2 during the cross-domain pre-training.

MiT with different Transformer scales. Transformer is the dominant component in the MiT backbone. We investigate the effect of Transformer scales in MiT. Specifically, we compare a MiT with 22 Transformer layers (MiT-22) and another with seven layers (MiT-7). The segmentation and classification performance is given in Figure 4, from which three conclusions can be drawn: (1) increasing the Transformer layers boosts the performance of MiT in all downstream tasks; (2) as MiT goes deeper, the performance gain of the dimension-specific pre-training over the random initialization becomes smaller, while the performance

Table 6. Segmentation and classification performance of our UniMiSS with different iteration intervals on the validation sets.

Iteration interval	BCV (3D seg)	RICORD (3D cls)	JSRT (2D seg)	ChestXR (2D cls)
1	82.70	82.95	92.33	96.65
2	**83.56**	**84.26**	**93.48**	**97.57**
3	83.28	83.65	93.12	97.16

Table 7. Segmentation performance of using the random initialization and three pre-training strategies on CHAOS dataset (unseen MRI scans) and ISIC dataset (unseen dermoscopic images).

Methods	Downstream data					
	2D dermoscopic			3D MRI		
	20%	40%	100%	20%	40%	100%
Rand. init.	76.31	79.92	85.07	73.28	83.64	88.38
MoCo v3	78.66	81.46	86.04	78.42	87.22	89.83
DINO	79.11	81.89	86.21	79.16	87.79	90.52
UniMiSS	**79.78**	**82.33**	**86.67**	**80.50**	**88.58**	**91.36**

gain of our UniMiSS with cross-dimension pre-training is basically impregnable; and (3) the superiority of our UniMiSS pre-training over the dimension-specific pre-training is more evident with the increase of Transformer layers.

Transferability on unseen modality data. In the above experiments, the pre-training and downstream tasks are all based on CT and X-ray images. To evaluate the transferability of UniMiSS on unseen modalities, we further tested the MoCo v3, DINO and our UniMiSS on the CHAOS dataset (MRI scans) and ISIC dataset (dermoscopic images). The results in Table 7 show that UniMiSS can consistently improve at least 2.98% on the CHAOS dataset, and 1.60% on the ISIC dataset, compared to the random initialization. It demonstrates that UniMiSS has a great potential in transferring learned knowledge to the unseen modality. Besides, our UniMiSS also outperforms two popular Transformer-based SSL methods on both CHAOS and ISIC datasets.

Necessity of SPE. Without the SPE module, a straightforward solution is to flatten the pixels or patches and then use a linear layer for the embedding. Such a crude flattening operation suffers the high computation complexity and memory requirements, especially for 3D images. Accordingly, SPE is an indispensable part of UniMiSS, which enables to (1) adaptively choose the patch embedding according to the input type; and (2) lessen the length of the sequence to reduce computation cost when the network goes deep.

Visualization of Segmentation Results. In Figure 5, we visualize the segmentation results obtained by the segmentation network, which is initialized (1) randomly, (2) by using the pre-trained MoCo v3 [13], (3) by using the pre-trained DINO [7], or (4) by using our pre-trained UniMiSS. It shows that our UniMiSS pre-training produces the higher-quality segmentation results, which are more similar to the ground truth, than MoCo v3 and DINO pre-training. Compared

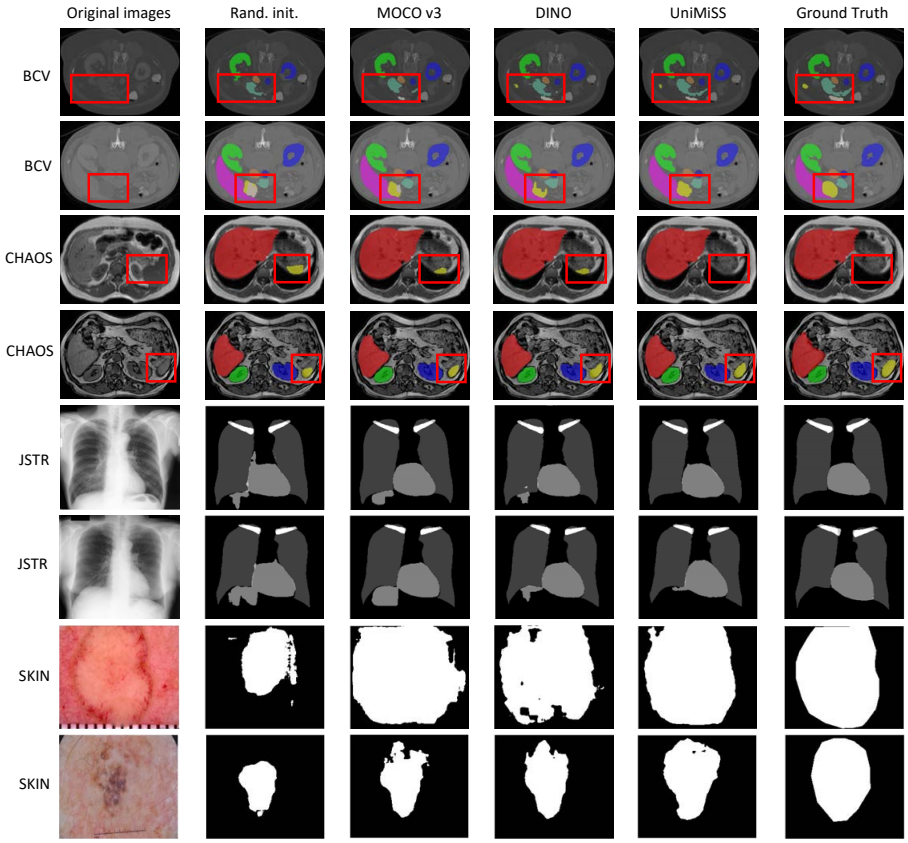

Fig. 5. Visualization of segmentation results of 8 cases selected from four datasets. The regions in red rectangles indicate our superiority. Our UniMiSS pre-training results in more accurate results than random initialization and other two pre-training strategies. Each type of organs and tumors in single dataset is denoted by a unique color.

to other competitors, UniMiSS pre-training is superior to process challenging cases, like small objects or blurry boundaries.

5 Conclusion

We propose a simple yet effective UniMiSS framework, which introduces a wealth of 2D medical images (*i.e.* X-rays) to the 3D SSL, aiming at making up for the lack of 3D data (*i.e.* CT scans). To break the difficulty of dimensionality barrier, we design the MiT as a bridge to connect different dimensions. In the future, we will extend our UniMiSS to deal with more dimensions (*e.g.* clinic text or genetic data).

Acknowledgement Jianpeng Zhang and Yong Xia were supported by National Natural Science Foundation of China under Grants 62171377. Qi Wu was funded by ARC DE190100539.

References

1. Multi-atlas labeling beyond the cranial vault - workshop and challenge. https://www.synapse.org/#!Synapse:syn3193805/wiki/217789 9
2. Tianchi dataset. https://tianchi.aliyun.com/competition/entrance/231601/information?from=oldUrl 8
3. Akhloufi, M.A., Chetoui, M.: Chest XR COVID-19 detection. https://cxr-covid19.grand-challenge.org/ (August 2021), online; accessed September 2021 9
4. An, P., Xu, S., Harmon, S., Turkbey, E., Sanford, T., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., et al.: Ct images in covid-19 [data set] https://doi.org/10.7937/TCIA.2020.GQRY-NC81. The Cancer Imaging Archive (2020) 8
5. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011) 8
6. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018) 3
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 4, 6, 7, 9, 10, 11, 13, 20
8. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: NeurIPS. vol. 33 (2020) 2, 4
9. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis **58**, 101539 (2019) 2, 4
10. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML. pp. 1691–1703 (2020) 3
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 3
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 9, 10, 11
13. Chen*, X., Xie*, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV (2021) 4, 6, 9, 10, 11, 13, 20
14. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: ISBI). pp. 168–172. IEEE (2018) 9
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 2, 5
16. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE Transactions on Medical Imaging **39**(7), 2415–2425 (2020) 4
17. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: NeurIPS (2020) 3

18. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022) 10, 11
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 3
20. 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. pp. 770–778 (2016) 11, 21, 23
21. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019) 3
22. Isensee, F., Jaeger, P.F., Kohl, S.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) 10, 11, 20
23. Jin, L., Yang, J., Kuang, K., Ni, B., Gao, Y., Sun, Y., Gao, P., Ma, W., Tan, M., Kang, H., Chen, J., Li, M.: Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. EBioMedicine (2020) 8
24. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain mr segmentation across scanners and protocols. In: MICCAI. pp. 476–484. Springer (2018) 4
25. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (Apr 2019). https://doi.org/10.5281/zenodo.3362844, https://doi.org/10.5281/zenodo.3362844 9
26. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR. pp. 6874–6883 (2017) 3
27. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. pp. 4681–4690 (2017) 3
28. Lee, H., Hwang, S.J., Shin, J.: Self-supervised label augmentation via input transformations. In: ICML (2020) 3
29. Li, K., Wang, S., Yu, L., Heng, P.A.: Dual-teacher++: Exploiting intra-domain and inter-domain knowledge with reliable transfer for cardiac segmentation. IEEE Transactions on Medical Imaging (2020) 4
30. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE Transactions on Medical Imaging **39**(9), 2713–2724 (2020) 4
31. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017) 9
32. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018) 9
33. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR. pp. 6707–6717 (2020) 3
34. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84. Springer (2016) 3
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3
36. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016) 3

37. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015) 3
38. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical Image Analysis **42**, 1–13 (2017) 2
39. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules http://db.jsrt.or.jp/eng.php. American Journal of Roentgenology **174**(1), 71–74 (2000) 9
40. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: Moco pretraining improves representation and transferability of chest x-ray models. In: MIDL. pp. 728–744. PMLR (2021) 2, 4
41. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. In: NeurIPS. vol. 33, pp. 18158–18172 (2020) 2, 4
42. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding (2020) 3
43. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: The rsna international covid-19 open radiology database (ricord). Radiology **299**(1), E204–E213 (2021) 9
44. Van Ginneken, B., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database https://www.isi.uu.nl/Research/Databases/SCR/index.php. Medical Image Analysis **10**(1), 19–40 (2006) 9
45. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV (2021) 5, 6, 20, 21
46. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 2097–2106 (2017) 8
47. Xie, Y., Zhang, J., Liao, Z., Xia, Y., Shen, C.: Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation. arXiv preprint arXiv:2011.12640 (2020) 2, 4, 9, 10, 11
48. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: MICCAI. pp. 171–180. Springer (2021) 2
49. Xie, Y., Zhang, J., Xia, Y., Shen, C.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Transactions on Medical Imaging **39**(7), 2482–2493 (2020) 2
50. Xie, Y., Zhang, J., Xia, Y., Shen, C.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Transactions on Medical Imaging **39**(7), 2482–2493 (2020) 20
51. Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., et al.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE Transactions on Medical Imaging **40**(3), 879–890 (2020) 2

52. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: CVPR. pp. 1195–1204 (2021) 2, 8, 10, 11
53. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: CVPR. pp. 1058–1067 (2017) 3
54. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: CVPR. pp. 9242–9251 (2018) 4
55. Zhou, H.Y., Lu, C., Yang, S., Han, X., Yu, Y.: Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In: ICCV. pp. 3499–3509 (2021) 2, 4, 9, 10, 11, 21
56. Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: ICCV. pp. 10672–10681 (2019) 10, 11
57. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021) 2, 4
58. Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y.: Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. Medical Image Analysis **64**, 101746 (2020) 2, 4

Appendix

A Overview

In this document, we provide more discussions and experimental details to supplement the main submission. We first continue to discuss the necessity of switchable patch embedding (SPE) module (Section B). We then give more details for the downstream tasks, including the implementation details and architectures (Section C). Finally, we provide an intuitive explanation of the proposed volume-slice consistency mechanism (Section D).

B Necessity of SPE (Cont.)

To further explain the necessity of the SPE module, we compared the pyramid U-like medical Transformer (MiT) to two variants without a SPE module. For the variant 1, we directly flatten the 2D/3D images to a sequence based on the pixels/voxels level and then use a linear layer for the embedding. Such a crude flattening operation suffers the very high computation complexity and memory requirements, especially for 3D images. Thus, it is hard to perform the variant 1 for quantitative comparisons. For the variant 2, we perform a naive embedding strategy to reduce the complexity. We first down-sample (for encoder)/up-sample (for decoder) the 2D/3D images by using a parameter free interpolation, then flatten them into a sequence based on the pixels/voxels level, and finally use a linear layer for both 2D and 3D embedding. The results in Table 8 show that MiT with the SPE module is significantly superior to the naive embedding strategy (*i.e.* variant 2) whenever with or without using the pre-training. It suggests that our SPE is better than the parameter-free interpolation and linear layer. The reason may be that the strided convolution with a large kernel is able to model the local continuity of 2D/3D images, which cannot be implemented by the linear layer.

Table 8. Segmentation performance of MiT and its two variants without a SPE module on BCV offline test set (3D CT).

Methods		SPE	Dice
Random initialization	Variant 1	No	unaffordable
	Variant 2	No	73.31
	Ours	Yes	79.93
UniMiSS pre-training	Variant 1	No	unaffordable
	Variant 2	No	76.65
	Ours	Yes	84.99

Table 9. Implementation details of downstream tasks. Seg: Segmentation; Cls: Classification; CE: Cross-entropy loss; off: offline test set; on: online test set.

Dataset	BCV	RICORD	JSRT	ChestXR	CHAOS	ISIC
Task	Seg	Cls	Seg	Cls	Seg	Seg
Modality	3D CT	3D CT	2D X-ray	2D X-ray	3D MRI	2D Dermoscopic
Training data	24	182	124	17,955	16	2000
Test data	6 (off)+20 (on)	45	123	3,430	4	600
Loss	Dice+CE [22]	CE	Dice+CE	CE	Dice+CE [22]	Hybrid loss [50]
Patch size	48×192^2	64×128^2	224^2	224^2	$48 \times 192 \times 256$	224^2
Augmentation	✓	✓	✓	✓	✓	✓
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	0.0001	0.00001	0.0001	0.0001	0.0001	0.0001
Batch size	2	8	32	32	2	16
Iterations	25,000	14,000	10,000	17,000	50,000	37,500

C Downstream Tasks

C.1 Implementation Details

In Table 9, we provide the implementation details of six downstream datasets, including the task type, modality, number of training and test cases, loss function, patch size, batch size, optimizer, learning rate, and maximum iterations. Note that we randomly split 25% training scans as a validation set to select the hyper-parameters of UniMiSS in the ablation study. We use the online data augmentation to alleviate the over-fitting of UniMiSS on training data. We augment 2D images via random cropping and zooming, random rotation, shear, shift, and horizontal/vertical flip. As for 3D images, we perform random rotation, scaling, flipping, adding white Gaussian noise, Gaussian blurring, adjusting rightness and contrast, simulation of low resolution, and Gamma transformation [22]. All the downstream experiments were performed on a NVIDIA GTX 2080Ti GPU.

C.2 Architectures of MiT and ResUnet

Figure 6 shows the detailed settings of the MiT network. The MiT encoder follows a progressive shrinking pyramid Transformer, as done in [45]. It consists explicitly of four stages, each of which is composed of a SPE module and several stacked Transformers. In each stage, the SPE module down-samples the input features and generates the dimension-specific embedded sequence. Notably, we append an extra learnable SSL token [7,13] to the patch embedded sequence. The SSL token is similar to the [CLS] token in ViT, which is able to aggregate information from the whole patch embedding tokens via the self-attention. The resultant sequences, combined with the learnable positional embedding, are inputted into the following Transformers for the long-term dependency modeling. Each Transformer layer includes a self-attention module and a feed-forward network (FFN) with two hidden layers. To reduce the computational cost and enable MiT to process high-resolution images, we follow the spatial-reduction

attention (SRA) layer to reduce the spatial complexity [45]. MiT has a symmetric decoder structure that consists of three stages. In each stage, the input feature map is first up-sampled by the SPE module, and then refined by the stacked Transformer layers. Besides, we also add skip connections between the encoder and decoder to keep more low-level but high-resolution information. We devise two MiT by changing the number of Transformer layers, namely MiT-7 and MiT-22. Noticed that default MiT-22 is used in the main submission unless otherwise specified.

Figure 7 shows the architecture of CNN-based ResUnet, used by the compared PCRL [55]. It consists of a 2D/3D ResNet-50 [20] encoder, a decoder, and four skip connections between encoder and decoder. The decoder contains five up-sampling modules. Each of the first four modules has a transposed convolutional (TransConv) layer followed by a convolution block (ConvBlock) and a pixel-wise summation with the corresponding feature maps from the encoder and the TransConv layer. The last module comprises an Up-sampling layer followed by a 1×1 Conv layer that maps each 32-channel feature map to the desired number of classes.

D Volume-slice consistency mechanism

Figure 8 gives an intuitive explanation of the proposed volume-slice consistency mechanism. Given a 3D volumetric image, we first create two augmented views via data augmentation, each of which has m 2D slices. We then compute the volumetric or slice representations of dual paths, *i.e.* $\boldsymbol{f}_1^{\text{Volume}}$, $\boldsymbol{f}_2^{\text{Volume}}$, $\boldsymbol{f}_1^{\text{Slices}}$, and $\boldsymbol{f}_2^{\text{Slices}}$. Here the slice representations $\boldsymbol{f}_1^{\text{Slices}}$ and $\boldsymbol{f}_2^{\text{Slices}}$ are generated by averaging the outputs of m slices. The loss function is composed of four items, including $\mathcal{L}^{\text{Volume}}$, $\mathcal{L}^{\text{Slices}}$, $\mathcal{L}^{\text{Volume}\rightarrow\text{Slices}}$, and $\mathcal{L}^{\text{Slices}\rightarrow\text{Volume}}$. The first two items aim to achieve the consistency at the level of global volume and local slices, respectively. Besides, the consistency across both levels should also be satisfied, which is achieved by the latter two items. By jointly using these four loss items, our model is able to capture richer representations from 3D medical images.

Layer_name			MiT-7		MiT-22		Output Size	
Encoder	SPE		2D	3D	2D	3D	2D	3D
			Kernel: 7×7 Channel: 32 Stride: 2	Kernel: 7×7×7 Channel: 32 Stride: (1, 2, 2)	Kernel: 7×7 Channel: 32 Stride: 2	Kernel: 7×7×7 Channel: 32 Stride: (1, 2, 2)	$\frac{H}{2} \times \frac{W}{2}$	$D \times \frac{H}{2} \times \frac{W}{2}$
	Stage 1	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 3×3 Channel: 48 Stride: 2	Kernel: 3×3×3 Channel: 48 Stride: 2	Kernel: 3×3 Channel: 48 Stride: 2	Kernel: 3×3×3 Channel: 48 Stride: 2	$\frac{H}{4} \times \frac{W}{4}$	$\frac{D}{2} \times \frac{H}{4} \times \frac{W}{4}$
		Transformer Layers	R = 6 H = 1 ×1 E = 4		R = 6 H = 1 ×2 E = 4		$\frac{H}{4} \times \frac{W}{4} + 1$	$\frac{D}{2} \times \frac{H}{4} \times \frac{W}{4} + 1$
	Stage 2	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 3×3 Channel: 128 Stride: 2	Kernel: 3×3×3 Channel: 128 Stride: 2	Kernel: 3×3 Channel: 128 Stride: 2	Kernel: 3×3×3 Channel: 128 Stride: 2	$\frac{H}{8} \times \frac{W}{8}$	$\frac{D}{4} \times \frac{H}{8} \times \frac{W}{8}$
		Transformer Layers	R = 4 H = 2 ×1 E = 4		R = 4 H = 2 ×3 E = 4		$\frac{H}{8} \times \frac{W}{8} + 1$	$\frac{D}{4} \times \frac{H}{8} \times \frac{W}{8} + 1$
	Stage 3	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 3×3 Channel: 256 Stride: 2	Kernel: 3×3×3 Channel: 256 Stride: 2	Kernel: 3×3 Channel: 256 Stride: 2	Kernel: 3×3×3 Channel: 256 Stride: 2	$\frac{H}{16} \times \frac{W}{16}$	$\frac{D}{8} \times \frac{H}{16} \times \frac{W}{16}$
		Transformer Layers	R = 2 H = 4 ×1 E = 4		R = 2 H = 4 ×4 E = 4		$\frac{H}{16} \times \frac{W}{16} + 1$	$\frac{D}{8} \times \frac{H}{16} \times \frac{W}{16} + 1$
	Stage 4	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 3×3 Channel: 512 Stride: 2	Kernel: 3×3×3 Channel: 512 Stride: 2	Kernel: 3×3 Channel: 512 Stride: 2	Kernel: 3×3×3 Channel: 512 Stride: 2	$\frac{H}{32} \times \frac{W}{32}$	$\frac{D}{16} \times \frac{H}{32} \times \frac{W}{32}$
		Transformer Layers	R = 1 H = 8 ×1 E = 4		R = 1 H = 8 ×3 E = 4		$\frac{H}{32} \times \frac{W}{32} + 1$	$\frac{D}{16} \times \frac{H}{32} \times \frac{W}{32} + 1$
Decoder	Stage 1	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 2×2 Channel: 256 Stride: 2	Kernel: 2×2×2 Channel: 256 Stride: 2	Kernel: 2×2 Channel: 256 Stride: 2	Kernel: 2×2×2 Channel: 256 Stride: 2	$\frac{H}{16} \times \frac{W}{16}$	$\frac{D}{8} \times \frac{H}{16} \times \frac{W}{16}$
		Transformer Layers	R = 2 H = 8 ×1 E = 4		R = 2 H = 8 ×3 E = 4		$\frac{H}{16} \times \frac{W}{16} + 1$	$\frac{D}{8} \times \frac{H}{16} \times \frac{W}{16} + 1$
	Stage 2	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 2×2 Channel: 128 Stride: 2	Kernel: 2×2×2 Channel: 128 Stride: 2	Kernel: 2×2 Channel: 128 Stride: 2	Kernel: 2×2×2 Channel: 128 Stride: 2	$\frac{H}{8} \times \frac{W}{8}$	$\frac{D}{4} \times \frac{H}{8} \times \frac{W}{8}$
		Transformer Layers	R = 4 H = 4 ×1 E = 4		R = 4 H = 4 ×4 E = 4		$\frac{H}{8} \times \frac{W}{8} + 1$	$\frac{D}{4} \times \frac{H}{8} \times \frac{W}{8} + 1$
	Stage 3	SPE	2D	3D	2D	3D	2D	3D
			Kernel: 2×2 Channel: 48 Stride: 2	Kernel: 2×2×2 Channel: 48 Stride: 2	Kernel: 2×2 Channel: 48 Stride: 2	Kernel: 2×2×2 Channel: 48 Stride: 2	$\frac{H}{4} \times \frac{W}{4}$	$\frac{D}{2} \times \frac{H}{4} \times \frac{W}{4}$
		Transformer Layers	R = 6 H = 2 ×1 E = 4		R = 6 H = 2 ×3 E = 4		$\frac{H}{4} \times \frac{W}{4} + 1$	$\frac{D}{2} \times \frac{H}{4} \times \frac{W}{4} + 1$

Fig. 6. Detailed settings of MiT network. Here, 'R': reduction ratio of SRA; 'H': head number of SRA; and 'E': expansion ratio of FFN

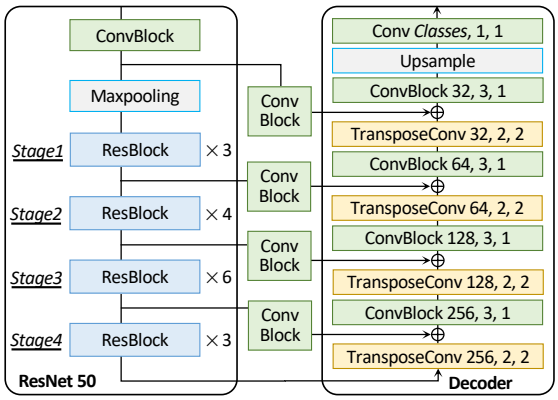

Fig. 7. Detailed architecture of ResUnet: A 2D/3D ResNet-50 [20] encoder, a decoder, and four skip connections between encoder and decoder. Green 'ConvBlock': 2D Conv-Batch Normalization(BN)-ReLU or 3D Conv-IN-LeakyReLU; Yellow 'TransConv': 2D/3D transposed convolutional layer. Note that the numbers in each block / layer indicate the number of filters, kernel size, and stride, respectively.

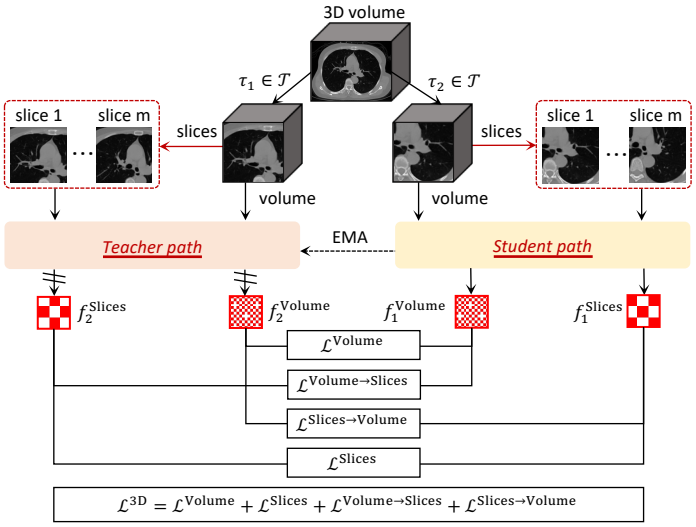

Fig. 8. Intuitive explanation of volume-slice consistency mechanism.