Unified 2D and 3D Pre-training for Medical Image Classification and Segmentation

Yutong Xie¹,², Jianpeng Zhang², Yong Xia², Qi Wu¹

¹ Australia Institute for Machine Learning, The University of Adelaide, Australia
² School of Computer Science and Engineering, Northwestern Polytechnical University, China

yutong.xie678@gmail.com; james.zhang@mail.nwpu.edu.cn; yxia@nwpu.edu.cn; qi.wu01@adelaide.edu.au

Abstract

Self-supervised learning (SSL) opens up huge opportunities for better utilizing unlabeled data. It is essential for medical image analysis that is generally known for its lack of annotations. However, when we attempt to use as many as possible unlabeled medical images in SSL, breaking the dimension barrier (i.e., making it possible to jointly use both 2D and 3D images) becomes a must. In this paper, we propose a Universal Self-Supervised Transformer (USST) framework based on the student-teacher paradigm, aiming to leverage a huge of unlabeled medical data with multiple dimensions to learn rich representations. To achieve this, we design a Pyramid Transformer U-Net (PTU) as the backbone, which is composed of switchable patch embedding (SPE) layers and Transformer layers. The SPE layer switches to either 2D or 3D patch embedding depending on the input dimension. After that, the images are converted to a sequence regardless of their original dimensions. The Transformer layer then models the long-term dependencies in a sequence-to-sequence manner, thus enabling USST to learn representations from both 2D and 3D images. USST has two obvious merits compared to current dimension-specific SSL: (1) more effective - can learn representations from more and diverse data; and (2) more versatile - can be transferred to various downstream tasks. The results show that USST provides promising results on six 2D/3D medical image classification and segmentation tasks, outperforming the supervised ImageNet pre-training and advanced SSL counterparts substantially.

1. Introduction

Medical imaging modalities can be categorized roughly into 2D imaging (e.g. X-ray) and 3D imaging (e.g. Computed Tomography (CT)). Limited by the dimension barrier, current self-supervised learning (SSL) approaches are designed for dimension-specific imaging [8,9,38,39,45,54–56], i.e. a SSL model can only be pre-trained on either 2D or 3D medical images and applied to the downstream tasks dealing with the same dimension images (see Figure 6, left panel). Few can handle both 2D and 3D medical images. This study aims to break the dimension barrier and develop a universal SSL model, which can not only be pre-trained on both 2D and 3D medical images but also perform well on various 2D or 3D downstream tasks (see Figure 6, right panel). Such a universal SSL model can make full use of large-scale pre-training, thus having at least two obvious and significant merits. First, it may achieve better performance on downstream tasks, since it boosts the model towards learning effective representations from more and diverse data. Second, it is a green computing technique, since the pre-trained model possesses impressive portability and can be directly transferred to various downstream medical image analysis tasks, avoiding pre-training different models for those tasks and thus decreasing carbon emissions.

However, building such a universal SSL model is challenging. Due to the dimension barrier, the common practice in medical image analysis is to design 2D convolutional neural networks (CNNs) for 2D images [47,50,54] and 3D CNNs for 3D images [45,46,51,54,55]. Although
2D networks can process 3D images on a slice-by-slice basis [8, 49], it may lose the inter-slice correlations and leads to poor performance. Considering its intrinsic structural limitation, it is almost impossible to design a universal SSL model for both 2D and 3D based entirely on convolutional operations.

Recent months have witnessed the success of Transformers in computer vision. A vision Transformer usually accepts a sequence of image patches, which was represented by learned linear embeddings, as an input to explore the relationships among the sequence elements. Once the embedded representations of the image patch sequence are determined, Transformer can accept the data of any dimensions, including but not limited to 2D images and 3D volumetric data, as its inputs. Therefore, Transformer offers the possibility of breaking the dimension barrier and constructing a universal SSL model.

In this paper, we propose a Universal Self-Supervised Transformer (USST) framework that is able to learn universal representations from both 2D and 3D unlabeled medical images. USST is based on the popular student-teacher paradigm. The student path directly predicts the output of the momentum-updated teacher. In each path, the backbone is a Pyramid Transformer U-Net (PTU) composed of switchable patch embedding (SPE) layers and Transformer layers. The SPE layer converts the input images to a sequence by using 2D or 3D patch embedding, depending on the input dimension. The Transformer layer then processes the embeddings in a sequence-to-sequence manner, regardless of their original dimension. The design of PTU enables USST to learn rich representations from both 2D and 3D images. Besides, we notice that the feature of a 3D volumetric image is homologous with the holistic features of the slices in this volume, and hence incorporate the slice-volume consistency into the loss function to make the most of 3D medical images.

To summarise, our contributions are two-fold.

1. We propose the USST framework, a simple and effective Transformer-based SSL approach, to learn general visual representations via pre-training on large-scale 2D and 3D medical images. The pre-trained USST is versatile and can perform well on six 2D/3D medical image classification and segmentation tasks, substantially beating the random initialization, supervised ImageNet pre-training, and advanced dimension-specific SSL opponents.

2. We consider the homology between the global features of a 3D volume and the local features of its slices, and thus design the slice-volume consistency mechanism to facilitate the model to learn rich visual representations from unlabeled 3D medical images.

2. Related Work

2.1. SSL approaches

CNN-based SSL approaches have been extensively studied in the literature. According to pretext tasks, these approaches can be broadly categorized into generative SSL approaches [25, 26, 35, 36, 52] and discriminative SSL approaches [6, 11, 17, 18, 20, 27, 32–34, 40]. Among them, contrastive learning methods [11, 17, 18, 20, 32, 34, 40] have drawn significant research attention and achieved advanced performance on many vision tasks. More recently, with the prevalence of Transformer in computer vision, vision Transformers have been introduced to SSL approaches. The seminal work is iGPT [10], which follows the masked auto-regressive language modelling paradigm to train self-supervised vision Transformers. Considering the success of contrastive learning in vision SSL, researchers attempted to train Transformers using the contrastive/Siamese paradigm in DINO [7] and MoCo v3 [13], which outperform a ResNet-based SSL approach on ImageNet and set a new record.

The success of SSL in computer vision has prompted investigators to apply SSL to medical image analysis [8, 9, 38, 39, 45, 54–56]. Typical solutions include pre-training a CNN by restoring the content of raw images [9, 39, 54–56] and tailoring contrastive SSL to medical images [8, 38, 39, 45]. These efforts constitute an important and timely step forward towards better SSL approaches to medical image analysis. However, they have two limitations. First, the CNN is pre-trained only on either 2D or 3D medical images, and hence its representation ability and portability would be trapped. Consequently, the pre-trained CNN can be transferred only to the downstream task dealing with the images of the same dimension. Second, current SSL approaches capture the spatial context of 3D medical images from either 2D slices [8] or the 3D volume [39, 55, 56]. Few of them jointly utilize both slices and volume information available in 3D images.

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 [23, 29] or using different imaging modalities [16, 28, 53] to train a single model that could perform well on diverse datasets. Karani et al. [23] and Liu et al. [29] 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. [53] simultaneously learned a volume-to-volume translation using unpaired CT and MRI data and strong segmentors using synthetic data, which were translated from another modality. Dou et al. [16] derived a new loss with knowledge


distillation (KD) to leverage the shared across-modality information between CT and MRI for accurate segmentation of anatomical structures. Li et al. [28] 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 3D medical images, which is different from our scenario. In this work, we aim to learn visual representations from more and diverse mix-up unlabeled 2D and 3D medical images, and thus provide a more general and effective solution for medical image analysis.

3. Methods

The objective of our USST is to use both 2D and 3D unlabeled medical images to learn general image representations, which are transferable and generalizable to downstream 2D and 3D medical image analysis tasks.

Figure 2 illustrates the pipeline of USST. Let an unlabeled training image sampled from the mix-up 2D and 3D data pool be denoted by $x^\#$, where $\#$ $\in \{2D, 3D\}$ represents the type of input data. The data augmentation module $\mathcal{T}$ draws two augmented views of $x^\#$, denoted by $x_1^\#$ and $x_2^\#$, which are then fed to the dual student-teacher path for feature representation learning. To adapt USST to both 2D and 3D medical images, we build a Pyramid Transformer U-Net (PTU) as the backbone, mainly constituted by SPE and Transformer layers. Each SPE layer switches to specific patch embedding depending on the dimension of the input image, whereas each Transformer layer is shared by both 2D and 3D inputs. Owing to those communal Transformer, USST can capture image information from either 2D or 3D data. We use 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 also introduce the slice-volume consistency loss, which encourages USST to jointly utilize both local slice and global volume information for training, thus is conducive to learning richer feature representations from volumetric images. We now delve into the details of this framework.

3.1. PTU Backbone

To make our self-supervised framework compatible with both 2D and 3D data and inclusive to different downstream scenarios, we constitute a PTU as the backbone, which incorporates the powerful Pyramid Vision Transformer (PVT) [43] into both the encoder and decoder of standard U-shaped architecture. Such a U-shaped design allows us to offer a pre-trained encoder for classification and a pre-trained encoder-decoder structure for segmentation. We now describe the PTU, and more details can be found in Appendix.

**SPE Layer.** The SPE layer is proposed to obtain the dimension-specific embedded sequence, which can switch to either 2D patch embedding or 3D patch embedding according to the dimension of the input image. Consequently, the SPE layer turns on 2D patch embedding when the input is a 2D image and turns on 3D patch embedding when the input is a 3D volume. Notice that the SPE in the encoder and decoder are different. The encoder SPE contains a switchable Conv-Instance Normalization(IN)-LeakyReLU block, which reduces the feature resolution. In contrast, the decoder SPE contains a switchable transpose convolutional layer, which increases the feature resolution.

**PTU Encoder.** The PTU encoder uses a progressive shrinking pyramid Transformer (i.e., PVT [43]). It consists explicitly of a SPE layer and four stages of Transformer blocks. Each Transformer block comprises a SPE layer and two to four Transformer layers (see Figure 2).

Each SPE layer is used to down-sample the input features and generate the dimension-specific embedded sequence. Notably, we append an extra learnable SSL token [7, 13] to the initial embedded sequence. The SSL token is similar to the class token in ViT, which is able to assemble knowledge from the other embeddings via self-attention. The resultant sequence added by learnable position embeddings is then encoded by the repeatedly stacking Transformer layers.

Each Transformer layer includes an attention layer and a feed-forward network (FFN) with two hidden layers. To reduce the computational cost and make PTU flexible to process high-resolution features, we use the spatial-reduction attention (SRA) layer [43]. Given a query $q$, a key $k$, and a value $v$ as the input, SRA first reduces the spatial resolution of $k$ and $v$, and then feeds $q$, reduced $k$, and reduced $v$ to the MSA layer to produce refined features. This process can be formally expressed as follows

$$SRA(q, k, v) = MSA(q, F(\sigma(R(k))), F(\sigma(R(v)))),$$

where $\sigma(\cdot)$ represents a linear projection that reduces the resolution of the input, $R(\cdot)$ resizes the input sequence to a feature map of the original size, and $F(\cdot)$ flattens the input into a 1D sequence.

With this architecture, the PTU encoder is able to learn high-resolution features at different scales. It hence can easier and better adapt to various downstream tasks, such as classification and segmentation, than ViT-like networks [15], which can only process low-resolution features.

**PTU Decoder.** The PTU decoder consists of three blocks. In each block, the input feature map is first up-sampled by a SPE layer, and then refined by three or four Transformer layers. Besides, we also add skip connections between the encoder and decoder to keep more low-level details. This decoder can achieve better decoding performance than a pure CNN decoder due to the strong ability of Transformer
Figure 2. Illustration of our proposed USST framework. It has a dual path (student and teacher) architecture. In each path, we use the PTU network as the backbone, mainly constituted by the SPE layers and Transformer layers. The SPE layer converts the input image (2D or 3D) to a sequence by using the either 2D or 3D patch embedding module. The Transformer layer then processes the embeddings in a sequence-to-sequence manner, thus enabling USST to learn representations from both 2D and 3D images. Note that the SPE is different in the encoder and decoder due to its specific role. The SPE in the encoder includes a switchable Conv-IN-LeakyReLU block to reduce the resolution. The SPE in the decoder includes a switchable transpose convolutional layer to increase the resolution.

3.2. USST Training

The proposed USST framework is based on the popular student-teacher paradigm. The student path comprises a PTU $\mathcal{F}^\#_\theta(\cdot)$ and a projector $P_\theta(\cdot)$, where $P_\theta(\cdot)$ is a $n$-layer multi-layer perceptron (MLP) head, $\theta$ is the ensemble of weights in this path, and $\# \in \{2D, 3D\}$. The SPE layers switch to the 2D patch embedding in $\mathcal{F}^{2D}_\theta(\cdot)$ and switch to the 3D patch embedding in $\mathcal{F}^{3D}_\theta(\cdot)$. Since the Transformer layer in PTU is able to break the dimension barrier, our USST can learn image representations from both 2D and 3D unlabeled medical images.

The teacher path has an identical architecture to that of the student path, but its weight $\mu$ updates with an exponential moving average (EMA) of the student path weights, shown as follows

$$\mu \leftarrow \lambda \mu + (1 - \lambda) \theta,$$

where $\lambda$ increases from 0.996 to 1 using a cosine schedule during training [7]. Since the teacher path just provides the ground truth signal for the student path, we apply a stop-gradient operator to the teacher.

In each path, we extract the SSL token after the last stage of PTU decoder, and feed it to the projector to produce the output of this path $f$, formally expressed as follows

$$f_1 = P_\omega(\mathcal{F}^\#_\theta(x_1^\#)), \quad f_2 = P_\mu(\mathcal{F}^\#_\mu(x_2^\#))$$

Our USST can be optimized to maximize the consistency between $f_1$ and $f_2$, which is calculated using different methods for 2D and 3D medical images, respectively. **Objective for 2D Images.** For 2D training images, we build a standard cross-entropy loss to minimize the discrepancy between the outputs of student and teacher paths $i.e.$, $f_1^{2D}$ and $f_2^{2D}$. To avoid model collapse, we also apply centering and sharpening operations [7] to the teacher outputs. The objective function can be formally presented as

$$\mathcal{L}^{2D}(x_1^{2D}, x_2^{2D}) = -\text{softmax}(\frac{f_2^{2D} - C}{\tau_t}) * \log(\text{softmax}(\frac{f_2^{2D}}{\tau_s}))$$

where $C$ is the center of teacher outputs, $\tau_t$ and $\tau_s$ are temperature parameters, and $\text{softmax}(\cdot)$ denotes the softmax function. The temperature $\tau_t$ is set to a small value in the teacher path for sharpening its outputs. The center $C$ is first computed with a mathematical expectation $E$ over the current teacher batch and then updated with an EMA to give consideration across different batches, shown as follows

$$C \leftarrow \omega * C + (1 - \omega) * E(f_2^{2D})$$

where $\omega$ is a rate parameter. Subtracting $C$ from $x_2^{2D}$ can hearten the model to the uniform distribution and has the op-
2D images 3D volumes 2D images 3D volumes

USST USST USST USST ...
2D-and-3D alternation training

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

Table 1. Downstream datasets for fine-tuning evaluation. "Num" denotes the number of images.

| Evaluation Datasets | Name | Tasks | Modalities | Num |
|---------------------|------|-------|------------|-----|
| JSRT [3,42]         | Multi-organ segmentation | 2D X-ray | 247 |
| ChestX [7]          | Pneumonia classification | 2D X-ray | \(21,390\) |
| BCV [1]             | Multi-organ segmentation | 3D CT | 30 |
| RICORD [41]         | COVID-19 screening | 3D CT | 227 |
| CHAOS [24]          | Abdominal organ segmentation | 3D MRI | 60 |
| ISIC [14]           | Skin lesion segmentation | 2D dermoscopy | 2750 |

Optimization Scheme. We introduce a 2D-and-3D alternative training scheme to solve this optimization problem. In each training step, we sequentially sampled batches of training data from the 2D or 3D medical image dataset for forward computation and error-backpropagation (see Figure 7). This alternative training process will not stop until our USST converges. The proposed training scheme has two advantages: (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, since the model is updated based on either 2D or 3D data in each \(v\) step.

4. Experiments

4.1. Datasets

Pre-training datasets. We collected 5,022 3D CT scans from five datasets (i.e., MOTS dataset [51], LIDC-IDRI dataset [5], Tianchi dataset [2], RibFrac dataset [22], TCI-ACT dataset [4]), and collected 108,948 2D images from NIH ChestX-ray8 dataset [44] to train our USST in a self-supervised manner.

Downstream datasets. Table 1 gives the downstream datasets that our pre-trained USST will be fine-tuned on, which cover six different tasks, including 2D multi-organ segmentation and pneumonia classification, 3D CT-based segmentation and classification, 3D MRI-based segmentation, and 2D skin lesion segmentation. Among them, the last two tasks are used to test the unseen-modality transferability. In practice, each dataset is split into the training set, the validation set and the test set, as shown in Appendix.
Table 2. Segmentation (seg) and classification (cls) performance of the random initialization (Rand. init.), dimension-specific pre-training (only 3D or 3D), and our proposed cross-dimension pre-training (USST).

| Methods         | 3D CT BCV (seg) | 3D CT RICORD (cls) | 2D X-ray JSRT (seg) | 2D X-ray ChestXR (cls) |
|-----------------|----------------|--------------------|---------------------|------------------------|
| Rand. init.     | 79.97          | 83.21              | 91.22               | 97.04                  |
| only 3D         | 83.73          | 87.26              | /                   | /                      |
| only 2D         | /              | /                  | 92.92               | 98.61                  |
| USST (2D+3D)    | 84.99          | 89.06              | 94.08               | 99.07                  |

4.2. Experimental Details

Pre-training setup. We set the size of input 2D patches to 224 × 224 and 3D patches to 16 × 96 × 96, aiming to weigh the balance between reserving enough information for SSL and reducing computational and spatial complexity to an affordable level. We applied data augmentations 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 [31] with a cosine decaying learning rate [30], a warm-up period of 10 epochs, to train our USST. We empirically set the initial learning rate to 0.0008, batch size to 192, maximum epochs 10 epochs, to train our USST. We empirically set the initial learning rate to 0.0008, batch size to 192, maximum epochs to 200, rate parameter $\omega$ to 0.9, and temperature parameter $\tau_1$ and $\tau_2$ to 0.04 and 0.1, respectively. It took about 2.5 days to pre-train the USST using 8 NVIDIA V100 GPUs.

Downstream training setup. For the classification, we extracted the pre-trained PTU 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 PTU encoder and decoder while removing the SSL token, and appended a segmentation head for prediction. This head includes a transposed convolutional layer, a 2D/3D ConvIN-LeakyReLU, and a 2D/3D 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). The detailed training setups for each downstream task are shown in Appendix.

4.3. Experimental Results

Dimension-specific vs. cross-dimension pre-training. Compared to the dimension-specific (2D or 3D) pre-training, our USST using the cross-dimension pre-training integrates multiple dimension images to perform the self-supervised learning in a unified framework. Such a rich set of images stands a chance of unleashing colossal potential in representation learning. To verify this, we compared our cross-dimension pre-training to the dimension-specific 2D or 3D pre-training. Note that all of these methods use the PTU backbone during the pre-training. As shown in Table 2, the cross-dimension pre-training can achieve higher performance than the dimension-specific pre-training, which proves the value of rich data in the self-supervised learning process. Moreover, the dimension-specific pre-training suffers from limited applications. For example, the 2D based pre-training can only be transferred to 2D downstream tasks, failing to work on 3D tasks. We can apply the pre-trained USST to any downstream tasks by breaking the barrier of dimensions, leading to superior portability.

To further demonstrate the superiority of our USST pre-training, we visualized four input patches with their attention maps obtained by the supervised method, dimension-specific pre-training, and our USST in Figure 8. Here, the supervised methods for visualization of CT and X-ray are trained on the BCV and JSRT datasets, respectively. It shows that, compared to the dimension-specific SSL, our USST can highlight more prosperous regions. Meanwhile, these regions have more similar semantic information to the supervised method.

Comparisons with advanced CNN-based pre-training. We compared our USST pre-training with advanced CNN-based pre-training strategies, including random initialization, ImageNet pre-training (INpre) (only for 2D tasks), Momentum Contrast v2 (MoCo v2), prior guided local model (PGL), and preservational contrastive representation learning (PCRL). Among them, INpre, MoCo v2 and PGL use the ResNet-50 as the backbone while PCRL uses the U-Net with the encoder of ResNet-50 as the backbone. All CNN-based SSL methods are pre-trained on the same 2D or 3D unlabeled medical images to ensure an unprejudiced comparison. The results in top panel of Table 3 and Table 4 show that (1) the SSL methods, PCRL, PGL and MoCo v2, can surpass INpre in 2D tasks, revealing that pre-training on a large-scale medical image dataset is more friendly to medical image tasks than pre-training on natural images; (2)
Table 3. Segmentation and classification performance of using different pre-training strategies on 2D tasks. For JSRT dataset, we compute the average Dice score of lung, heart and clavicle. The ↑ represents the performance gain of pre-training strategies over random initialization.

| Methods         | SSL backbone     | JSRT (X-ray, seg) | ChestXR (X-ray, cls) |
|-----------------|------------------|-------------------|----------------------|
| Rand. init.     | /                | 20%   | 40%   | 100%  | 20% | 40% | 100%|
| INpre [19]      | CNN-based SSL    | 82.57 | 85.55 | 82.73 | 78.16 | 89.01 | 85.91 | 73.76 | 85.46 | 97.54 | 91.39 | 97.09 | 82.96 | 78.88 |
| MoCo v2 [12]    | 2D ResNet-50     | 82.57 | 85.55 | 82.73 | 78.16 | 89.01 | 85.91 | 73.76 | 85.46 | 97.54 | 91.39 | 97.09 | 82.96 | 78.88 |
| MoCo v3 [13]    | Transformer-based SSL | 82.57 | 85.55 | 82.73 | 78.16 | 89.01 | 85.91 | 73.76 | 85.46 | 97.54 | 91.39 | 97.09 | 82.96 | 78.88 |
| MoCo v3 [13]    | PTU              | 96.61 | 76.06 | 87.90 | 77.81 | 74.54 | 95.20 | 70.09 | 98.13 | 100% | 92.80 | 98.99 | 98.59 | 77.96 |
| DINO [7]        | PTU              | 96.61 | 76.06 | 87.90 | 77.81 | 74.54 | 95.20 | 70.09 | 98.13 | 100% | 92.80 | 98.99 | 98.59 | 77.96 |
| USST (Ours)     | PTU              | 91.88 | 93.15 | 94.08 | 94.08 | 92.86 | 97.09 | 94.29 | 98.14 | 92.94 | 99.07 | 92.03 | 77.96 | 94.08 |

Table 4. Segmentation and classification performance of using different pre-training strategies on 3D tasks. For BCV dataset, we compute the average Dice score of 13 organs.

| Methods         | SSL backbone     | BCV (CT, seg) | RICORD (CT, cls) |
|-----------------|------------------|---------------|------------------|
| Rand. init.     | /                | 20% | 40% | 100% | 20% | 40% | 100%|
| MoCo v2 [12]    | CNN-based SSL    | 68.44 | 73.14 | 79.93 | 69.72 | 74.66 | 83.36 |
| MoCo v3 [13]    | 3D ResNet-50     | 71.22 | 75.09 | 75.79 | 71.22 | 75.09 | 75.79 | 71.22 | 75.09 | 75.79 | 71.22 | 75.09 | 75.79 |
| PGL [54]        | 3D ResNet-50     | 72.05 | 75.86 | 75.86 | 72.05 | 75.86 | 75.86 | 72.05 | 75.86 | 75.86 | 72.05 | 75.86 | 75.86 |
| PCRL [54]       | 3D Res50Unet     | 72.80 | 76.05 | 76.05 | 72.80 | 76.05 | 76.05 | 72.80 | 76.05 | 76.05 | 72.80 | 76.05 | 76.05 |
| Rand. init.     | /                | 70.09 | 74.60 | 79.97 | 71.36 | 76.06 | 83.21 |
| MoCo v3 [13]    | Transformer-based SSL | 74.54 | 78.16 | 78.16 | 74.54 | 78.16 | 78.16 | 74.54 | 78.16 | 78.16 | 74.54 | 78.16 | 78.16 |
| DINO [7]        | PTU              | 75.33 | 78.88 | 78.88 | 75.33 | 78.88 | 78.88 | 75.33 | 78.88 | 78.88 | 75.33 | 78.88 | 78.88 |
| USST (Ours)     | PTU              | 77.96 | 80.97 | 84.99 | 77.96 | 80.97 | 84.99 | 77.96 | 80.97 | 84.99 | 77.96 | 80.97 | 84.99 |

in the case of random initialization, PTU is almost on par with, or slightly superior to the ResNet-50 network on four different tasks; and (3) compared to CNN-based SSL methods, our USST achieves a higher performance gain over random initialization. The superior performance corroborates the potential of USST in learning effective representations from unlabeled medical images.

Comparisons with advanced Transformer-based SSL. We also compared our USST against two Transformer-based SSL methods (i.e., MoCo v3 and DINO). Here MoCo v3, DINO and our USST use the PTU as the backbone network. Note that MoCo v3 and DINO are only pre-trained on either 2D or 3D medical images, while our USST can work on both 2D and 3D medical images. The results in bottom panel of Table 3 and Table 4 show that MoCo v3 and DINO can boost the performance apparently when compared to the random initialization. Our USST outperforms MoCo v3 and DINO by +1.05% on 2D segmentation, +0.37% on 2D classification, +2.38% on 3D segmentation and +3.15% on 3D classification, when 100% labels are available. It suggests that our pre-trained USST contributes to better and more robust performance than other Transformer-based SSL methods on the downstream medical image analysis tasks.

Results with fewer downstream annotations. The major advantage of our USST is to utilize the knowledge learned from cheap unlabeled images to facilitate downstream tasks, leading to (1) the less requirement of annotations and (2) improved performance, particularly when the annotated downstream training dataset is small. To demonstrate this advantage, we kept the test set and pre-trained weights unchanged, and randomly sampled 20% and 40% data from each training set, respectively, to fine-tune the downstream network. From Table 3 and Table 4, we can easily find that (1) the pre-trained USST beats steadily the random initialization and other pre-trained SSL approaches, no matter how many ratios of annotated data are used for fine-tuning, and leads to a more significant performance gain when the annotated training dataset is smaller; and (2) the downstream network initialized by the pre-trained USST and fine-tuned with 40% annotated images can achieve a comparable even better performance than that trained from scratch with 100% annotated images. A similar phenomenon can be observed when the number of annotated training images reduces to 40% and 20%, respectively. The results suggest that our USST provides the possibility of using unlabeled data to replace almost half of the labelled training images while maintaining the segmentation/classification performance.

4.4. Ablation

Effectiveness of slice-volume consistency. We design the slice-volume consistency mechanism for learning more rich
Table 5. Segmentation and classification performance on two 3D validation sets with or without using slice-volume consistency.

| Volume | Slices | BCV (seg) | RICORD (cls) |
|--------|--------|-----------|--------------|
| 50%    | 40%    | 72.08     | 76.04        |
| 100%   | 20%    | 90.17     | 95.14        |

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

| Iteration interval | 3D CT BCV (seg) | 3D CT RICORD (cls) | 2D X-ray JSRT (seg) | 2D X-ray ChestXR (cls) |
|--------------------|---------------|-------------------|-------------------|----------------------|
| 0                  | 78.55         | 78.13             | 90.17             | 95.14                |
| 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                |

4.5. Discussions

PTU with different Transformer scales. Transformer is the dominant component in the PTU backbone. We investigate the effect of Transformer scales in PTU. Specifically, we compare a PTU with 22 Transformer layers (PTU-22) and another with seven layers (PTU-7). The segmentation and classification performance is given in Figure 5, from which three conclusions can be drawn: (1) increasing the Transformer layers boosts the performance of PTU in all downstream tasks; (2) as PTU goes deeper, the performance gain of the dimension-specific pre-training over the random initialization becomes smaller, while the performance gain of our USST with cross-dimension pre-training is basically impregnable; and (3) the superiority of our USST 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 based on CT and X-ray images. To evaluate the transferability of USST on unseen modality images, we further tested the network initialized with the pre-trained MoCo v3, DINO and our USST on the CHAOS dataset (MRI scans) and ISIC dataset (dermoscopic images) in different ratios. The results in Table 7 show that USST can consistently improve the random initialization at least by +2.98% on the CHAOS dataset and by +1.60% on the ISIC dataset, which demonstrates that USST has a great potential in transferring learned knowledge to the unseen modality. Besides, our USST also outperforms two popular Transformer-based SSL methods on both CHAOS and ISIC datasets, which again verifies the superiority of our USST.

5. Conclusion

This paper has developed a simple and effective USST framework to break the barrier between 2D and 3D pre-training in medical images. USST introduces the Transformer as a bridge to connect different dimensions, thus enabling learning more universal representations from both 2D and 3D medical images and being easily transferred to various downstream applications. Experimental results on six different 2D/3D medical vision tasks show that
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). For CHAOS dataset, we compute the average Dice score of liver, kidneys and spleen.

| 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 |
| USST        | 79.78 82.33 86.67 | 80.50 88.58 91.36 |

USST significantly outperforms the supervised ImageNet pre-training and advanced dimension-specific SSL opponents. As a universal framework, our USST can be extended to deal with more dimensions (e.g. clinic text or genetic data) in the future.

References

[1] Multi-atlas labeling beyond the cranial vault - workshop and challenge. https://www.synapsea.org/#!Synapsea:sys3193805/wiki/217789
[2] Tianchi dataset. https://tianchi.aliyun.com/competition/entrance/231601/information?from=oldUrl
[3] Moulay A. Akhloufi and Mohamed Chetouii. Chest XR COVID-19 detection. https://cxr-covid19.grand-challenge.org/, August 2021. Online; accessed September 2021.
[4] P An, S Xu, S Harmon, E Turkbey, T Sanford, A Amalou, M Kassin, N Varble, M Blain, V Anderson, et al. Ct images in covid-19 [data set] https://doi.org/10.7937/TClA.2020.GQRY-NC81. The Cancer Imaging Archive, 2020.
[5] Samuel G Armato III, Geoffrey McNennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binh engage Zhao, Denise R Aberle, Claudia H Henschke, Eric A Hoffman, et al. The lung image database consortium (lidi) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics, 38(2):915–931, 2011.
[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In ECCV, pages 132–149, 2018.
[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021.
[8] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In NeurIPS, volume 33, 2020.
[9] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. Medical Image Analysis, 58:101539, 2019.
[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In ICML, pages 1691–1703, 2020.
[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.
[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.
[13] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In ICCV, 2021.
[14] Noel CF Codella, David Gutman, Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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, pages 168–172. IEEE, 2018.
[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.
[16] Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. IEEE Transactions on Medical Imaging, 39(7):2415–2425, 2020.
[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Ghaslaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In NeurIPS, 2020.
[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, pages 9729–9738, 2020.
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–777, 2016.
[20] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Michael F McNitt-Gray, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In ICLR, 2019.
[21] 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.
[22] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, Jiajun Chen, and Ming Li. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. EBioMedicine, 2020.
[23] Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A lifelong learning approach to brain localization and activity identification. In NeurIPS, volume 33, 2020.
mr segmentation across scanners and protocols. In *MICCAI*, pages 476–484. Springer, 2018. 2

[24] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. Apr. 2019. 5

[25] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, pages 6874–6883, 2017. 2

[26] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 2

[27] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Self-supervised label augmentation via input transformations. In *ICML*, 2020. 2

[28] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 39(9):2713–2724, 2020. 2

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2018. 6

[30] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, pages 171–180. Springer, 2021. 1

[31] Meendi Noroozi and Paolo Favaro. Unsupervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[32] Qian Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020. 2

[33] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[34] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2534–2544, 2016. 2

[37] Alex Radford, Luke Metz, and Soumith Chintala. Unsupervised learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[38] Junji Shirai, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitsue Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. 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. 5

[39] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *MIDL*, pages 728–744. PMLR, 2021. 1, 2

[40] Jianpeng Zhang, Yutong Xie, Jingbo Yang, Andrew Y. Ng, and Michael Maire. PGL: Prior-guided self-supervised learning for 3D medical image segmentation. *arXiv preprint arXiv:2111.12640*, 2020. 1, 2, 7

[41] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *CVPR*, pages 8280–8289, 2018. 2

[42] Jiheon Ye, Jiyoung Yu, Sangyoon Lee, and Joonseok Lee. Hierarchical convnet for medical image segmentation. *arXiv preprint arXiv:2010.00532*, 2020. 2

[43] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, pages 1058–1067, 2017. 2
A. Appendix

In the appendix, we provide more details and qualitative results to supplement the main submission. We first introduce more details for the downstream tasks, including the implementation details and architectures (Section B). We then provide an intuitive explanation of the proposed slice-volume consistency mechanism (Section C). Finally, we visualize the segmentation results of 8 cases randomly selected from four datasets for a qualitative comparison (Section D).

B. Downstream Tasks

B.1. Implementation Details

In Table 8, 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 USST in the ablation study. We use the online data augmentation to alleviate the over-fitting of USST 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 [21]. All the downstream experiments were performed on a NVIDIA GTX 2080Ti GPU.

B.2. Architectures of PTU

Figure 6 shows the detailed settings of the PTU network. It consists of an encoder, a decoder, and three skip connections between the encoder and decoder. The PTU encoder follows the progressive shrinking pyramid Transformer-like PVT [43], which contains a switchable patch embedding (SPE) layer and four stages of Transformer blocks. Each Transformer block is comprised of a SPE layer and several Transformer layers. The PTU decoder progressively up-samples the features to the desired resolution via the SPE layer and then refines the up-sampled features using several Transformer layers. Note that the SPE layer in the encoder is the switchable Conv-Instance Normalization(IN)-LeakyReLU block, which is used to obtain the dimension-specific embedded sequence and reduce the feature resolution, whereas the SPE layer in the decoder is the switchable transpose convolutional layer, which is used to up-sample the dimension-specific features and improve their resolution.
Table 8. Implementation details of downstream tasks. Dice: Dice loss; CE: Cross-entropy loss.

| Dataset | Task | BCV | RICORD | JSRT | ChestXR | CHAOS | ISIC |
|---------|------|-----|--------|------|---------|-------|------|
| Modality | Segmentation | 3D CT | 3D CT | Segmentation | Classification | Segmentation | Segmentation |
| Training data | 24 | 182 | 124 | 17,955 | 16 | 2000 |
| Test data | 6 | 45 | 123 | 3,430 | 4 | 600 |
| Loss | Dice+CE [21] | CE | Dice+CE | CE | Dice+CE [21] | Hybrid loss [48] |
| Patch size | 48 x 192 x 192 | 64 x 128 x 128 | 224 x 224 | 224 x 224 | 48 x 192 x 256 | 224 x 224 |
| Online augmentation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Optimizer | AdamW | AdamW | AdamW | AdamW | AdamW | AdamW |
| Learning rate | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| Batch size | 2 | 8 | 32 | 32 | 2 | 16 |
| Max. iterations | 25,000 | 14,000 | 10,000 | 17,000 | 50,000 | 37,500 |

![Layer name](https://example.com/layer.png)

Figure 6. Detailed settings of PTU network. Here, ‘R’: reduction ratio of SRA; ‘H’: head number of SRA; and ‘E’: expansion ratio of FFN

Figure 7. Intuitive explanation of slice-volume consistency mechanism.

**C. Slice-volume consistency mechanism**

Figure 7 gives an intuitive explanation of the proposed slice-volume 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., \( f_1^\text{Volume}, f_2^\text{Volume}, f_1^\text{Slices}, \) and \( f_2^\text{Slices} \). Here the slice representations \( f_1^\text{Slices} \) and \( f_2^\text{Slices} \) are generated by averaging the outputs of m slices. The loss function is composed of four items, including \( L^\text{Volume}, L^\text{Slices}, L^\text{Volume→Slices}, \) and \( L^\text{Slices→Volume} \). 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.

**D. Visualization of Segmentation Results**

In Figure 8, we visualize the segmentation results obtained by the segmentation network, which is initialized (1)
Figure 8. Visualization of segmentation results of 8 cases randomly selected from four datasets. The regions in red rectangles indicate our superiority. Our USST 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.

randomly, (2) using the pre-trained MoCo v3 [13], (3) using the pre-trained DINO [7], or (4) using our pre-trained USST. It shows that our USST pre-training produces the higher-quality segmentation results, which are more similar to the ground truth, than MoCo v3 and DINO pre-training. Compared to other competitors, USST pre-training is superior to process challenging cases, like small objects or blurry boundaries.