Self-Supervised Multi-Task Representation Learning for Sequential Medical Images

Anonymous

Anonymous Organization
@****.***

Abstract. Self-supervised representation learning has achieved promising results for downstream visual tasks in natural images. However, its use in the medical domain, where there is an underlying human structural similarity, remains underexplored. To address this shortcoming, we propose a self-supervised multi-task representation learning framework for sequential 2D medical images, which explicitly aims to exploit the underlying structures via multiple pretext tasks. Unlike the current state-of-the-art methods, which are designed to only pre-train the encoder for instance discrimination tasks, the proposed framework can pre-train the encoder and the decoder at the same time for dense prediction tasks. We evaluate the representations extracted by the proposed framework on two public whole heart segmentation datasets from different domains. The experimental results show that our proposed framework outperforms MoCo V2, a strong representation learning baseline. Given only a small amount of labeled data, the segmentation networks pre-trained by the proposed framework on unlabeled data can achieve better results than their counterparts trained by standard supervised approaches.

Keywords: Self-Supervised Learning · Multi-Task Learning · Medical Image Segmentation.

1 Introduction

Fueled by recent success of convolutional neural networks (CNN), deep learning (DL) has led many breakthroughs in computer vision tasks, benefiting from large-scale training data. However, under standard supervised learning (SL), preparing a large training dataset requires extensive and costly human annotation, especially in medical domain, where the annotation requires further domain expertise from the clinicians. To mitigate this data scarcity challenge in SL, there is a renaissance of research on self-supervised learning (SSL) [33]. SSL aims to learn meaningful representations from the unlabeled data in SSL and then transfer the extracted representations for the downstream task with a small-scale labeled data . For simplicity, we use the term self-supervised learning and the term self-supervised representation learning interchangeably in this work. The state-of-the-art (SOTA) SSL methods [19, 26, 6, 35, 8] have demonstrated that a model trained with only unlabeled data plus a small amount of labeled data can achieve comparable performance on various downstream tasks with the same model trained with a large amount of labeled data.

Although similar SSL techniques have been applied in medical tasks and have achieved promising performance, several questions remain inconclusive. First, SOTA SSL methods leverage instance-wise differences among the unlabeled data by *contrastive learning*. For example, the models are pre-trained on ImageNet [10] and fine-tune with PASCAL VOC [14] and MS COCO [23] for the downstream tasks. As shown in Fig. 1, compared with natural images from ImageNet [10] which belong to particular categories, the medical images show clear difference: (1) a natural image usually has the single object of interest centered in the image but a medical image usually contain more than one semantic class; (2) the background (BG) in medical images usually contains more supportive semantic information than natural images; (3) for a particular class, the instance-wise difference is more obvious in natural images than medical images. For medical images such as computerised tomography (CT) and magnetic resonance imaging (MRI), a scan contains a series of slices for the same patient. The difference between neighbored slices is commonly negligible by human eyes (see Fig. 2 for example). Second, to utilize instance discrimination as the *pretext* task, SOTA SSL methods can only pre-train the encoder for image classification tasks, due to the nature of instance discrimination task. In contrast to image classification, where an input image is mapped to a single label, dense prediction tasks are expected to learn a pixel-wise mapping between the input and the output. For downstream tasks such as depth estimation, edge detection, and surface normal estimation, contrastive learning cannot learn representations for the decoder. Third, medical datasets are usually much smaller than general-purpose datasets such as ImageNet. This would limit the performance of contrastive learning methods, which rely on large-scale training data to catch the instance-wise difference [19, 6]. Last but not least, directly applying SSL methods on medical images does not utilize domain-specific knowledge that is particular to the medical domain. Unlike general objects, medical objects such as human organs or human structures usually share statistical similarities in terms of the location, shape, and size among different patients [13]. There have been studies of utilizing such a human structural similarity, as a free lunch, in medical image analysis [9, 12, 13, 29]. See Fig. 2 for the intuition of human structural similarity.

To bridge the methodological gaps discussed above, we propose a novel SSL framework for medical images such as CT scans or MRI scans. We propose two pretext tasks that can be formulated as two SSL problems by utilizing the characteristics of the medical data. Concretely, for a single slice in a series of medical images in a sequential order, we try to reconstruct two slices that have a fixed distance to it on both sides. Given a CNN with an end-to-end pixel-wise mapping (e.g. U-Net [32]), two pretext tasks are trained jointly in a multi-task learning (MTL) formulation to learn domain-specific knowledge for the CNN. Given limited data, we use MTL to improve the generalization of the CNN. If a CNN can be decomposed into an encoder and decoder separately (e.g. FCN [24] where the encoder can be viewed as a standard feature extractor for im-

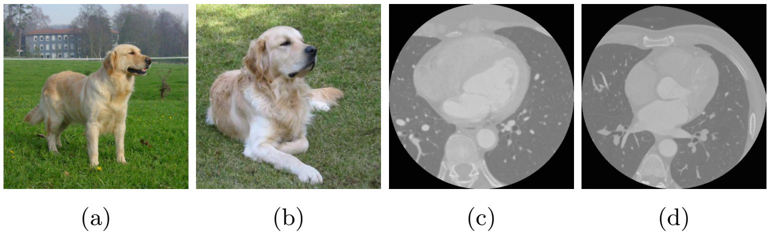


Fig. 1: Visual Comparison between instance-wise difference. (a) and (b) are two Golden Retrievers sampled from the ImageNet dataset [10]. (c) and (d) are two slices sampled from two patients' CT scans, which are collected in the CT-WHS dataset [42].

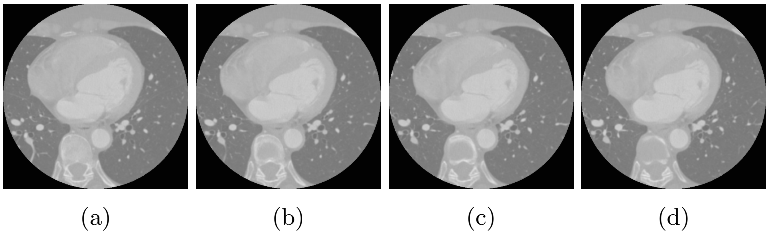


Fig. 2: A piece of sequential slices of a patient's CT scan. The CT scan is collected in the CT-WHS dataset [42].

age classification), we extend the proposed framework to integrate the instance discrimination task for the encoder as the third pretext task. We evaluate the proposed framework in medical image segmentation tasks where we pre-train a segmentation network on unlabeled data first and then fine-tune the model with a small-scale labeled data. The segmentation network pre-trained by the proposed framework can outperform the same network pre-trained by MoCo [19], the SOTA SSL method, in the whole heart segmentation tasks.

Our main contributions can be summarized as follows:

1. We propose a simple self-supervised learning framework for dense prediction tasks on medical images with inherent sequential order.
2. We are the first to exploit human structural similarity to integrate multi-task learning with self-supervised learning.
3. We extend the proposed framework to incorporate the concept of contrastive learning.

2 Related Works

2.1 Self-Supervised Learning

First formulated in [33], self-supervised learning (SSL) is a form of unsupervised learning where the learning process is not supervised by human-annotated

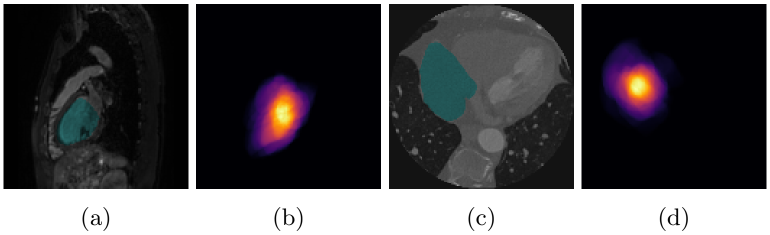
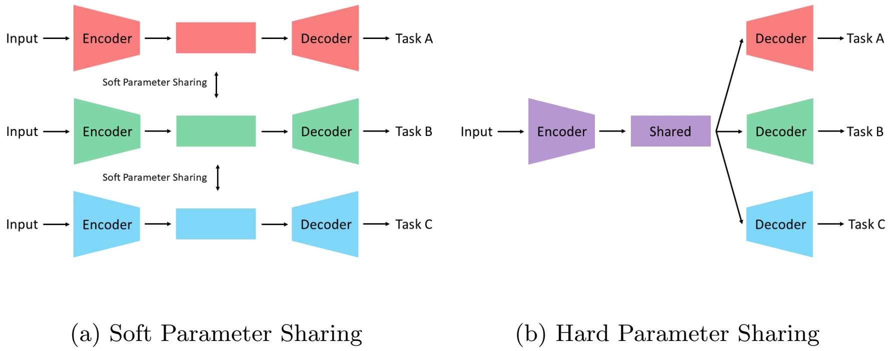


Fig. 3: Illustration of human structural similarity: (a) An axial CT image with the ground truth annotation of the right atrium. (b) The label distribution (normalized density heatmap) of right atriums in the CT-WHS dataset [41]. (c) A sagittal MRI image with the ground truth annotation of the left ventricle. (d) The label distributions of left ventricles in the MRI-WHS dataset [42].

labels. By observing how babies interacting with the new environment [17], the concept of SSL originated from cognitive science. Meanwhile, representation learning aims to extract useful representations from data [2] in the context of DL. An important application of SSL is to learn transferable representations for downstream tasks, e.g. common downstream tasks for visual understanding include image recognition [34, 20], object detection [31, 22], and semantic segmentation [24, 5]. With this definition, SSL can also be understood as transfer learning from unlabeled data [30].

Pretext Tasks In the recent renaissance of SSL in visual understanding tasks, the role of SSL in each taget task is associated with the corresponding *pretext* task. By solving a pretext task, the model extracts meaningful representations for the target task. That is to say, the model is pre-trained on the pretext task for the target task. For example, [11] utilizes two CNNs with shared weights to predict the relative positions of two patches randomly cropped from the same image. Similarly, for an image divided into a 3×3 grid, [27] permutes the order of 9 patches and predicted the index of the chosen permutation, just like solving jigsaw puzzles.[39] colorizes the grayscale images by using the lightness channel L as input to predict the corresponding a and b color channels of the image in the CIE *Lab* colorspace. [16] randomly rotates the images by multiples of 90 degrees and predicts the rotation. Designing a good pretext tasks requires extra effort and is challenging, as *a good self-supervised task is neither simple nor ambiguous* [27].

Contrastive Learning Contrastive learning was first developed as a learning paradigm for neural networks to identify what makes two objects similar or different [1]. In the literature of SSL, contrastive learning utilizes the instance-wise difference and systematically defines the pretext tasks as a simple instance discrimination task [19]. Recently, state-of-the-art (SOTA) contrastive learning methods [6, 19, 26, 35] have been proposed based on a contrastive loss, InfoNCE



(a) Soft Parameter Sharing (b) Hard Parameter Sharing

Fig. 4: Two common MTL workflows for dense prediction tasks given the same input. (a) Soft parameter sharing: The different tasks have separate models (in different colors), where the parameters are communicated between models. (b) Hard parameter sharing: The different tasks share the same encoder and network backbone (in purple) but independent decoders.

[28], which is motivated by *noise-contrastive estimation* (NCE) [18]. By minimizing InfoNCE, the model is expected to learn the invariant features shared by a positive pair [26, 6, 35], where a positive pair is usually defined as two sarcastically augmented views from the same instance. Note, SOTA contrastive learning methods usually rely on large-scale datasets, which are often unavailable in the medical domain.

2.2 Multi-Task Learning

Multi-task learning (MTL) [4] is a learning paradigm inspired by human learning activities where the knowledge learned from previous tasks can help learn a new task. MTL aims to improve the generalization performance of all the tasks by leveraging useful information contained in multiple related tasks [37]. In the era of DL, we use a model to map the input to the output, given a specific task. In contrast to single-task learning, where each task is handled by an independent model, MTL can reduce the memory footprint, increase overall inference speed, and improve the model performance. Moreover, when the associated tasks contain complementary information, MTL can regularize each single task. For dense prediction tasks, a good example is semantic segmentation, where we always assume that the classes of interest are mutually exclusive. Depending on the data modality of the input and the task affinity [36] between tasks, there are various types of MTL. We depict the workflows for the situations that the tasks share the same input in Fig. 4. Given the same input, pixel-level tasks in visual understanding often have similar characteristics, which can be potentially used to boost the performance by MTL [40].

3 Method

3.1 Problem Formulation

Let X be an unlabeled set consisting of sequences of medical images for N patients, i.e. $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$. Each example consists of a sequence of medical images $\boldsymbol{x}_i = \{x_i^j\}_{j=1}^{n_i}$ for the patient i. For example, the sequences of medical images could be CT scans or MRI scans. The goal is to learn meaningful representations from X for a downstream dense prediction task, such as semantic segmentation on medical images.

3.2 Self-Supervised Multi-Tasking Learning

Based on the empirical observation of human structural similarity, we assume that each medical image x_i^j follows an unknown distribution $\mathcal{X}$. We aim to utilize this similarity. Given a neural network backbone for a dense prediction task, we create two pretext tasks which can be formulated in a MTL setting. Concretely, given a sequence $\boldsymbol{x}_i$, we use an anchor image x_i^j to predict the image t steps before x_i^{j-t} and the image t steps behind x_i^{j+t}, where t is an integer. Formally, let the neural network backbone we are interested (including the encoder and the decoder) be f_θ and the auxiliary task decoders be g_{ϕ^-} for x_i^{j-t} and g_{ϕ^+} for x_i^{j+t} respectively. The loss function is

$$\mathcal{L}_{pretext} = \sum_i ||g_{\phi^-}(f_\theta(x_i^j)) - x_i^{j-t}|| + ||g_{\phi^+}(f_\theta(x_i^j)) - x_i^{j+t}|| \tag{1}$$

, where $|| \cdot ||$ denotes a distance measure in Euclidean space. For simplicity, we use a standard Euclidean distance (i.e. mean squared error). The overall learning framework is illustrated in Fig. 5.

Intuitively, when we are learning the mappings from x_i^j to x_i^{j-t} and x_i^{j+t}, the human structural similarity (e.g. the relative location, shape, and size of the organs and structures) is extracted by the neural network backbone. For the downstream tasks such as medical image segmentation, the extracted knowledge should play an important role as there is an overlap of the semantic information shared between the pretext tasks and the downstream tasks. Note, without the regularization of MTL, i.e. if there is only one pretext task, the neural network backbone could only memorize information for just one direction, which might be a easy pretext task to learn and make the learned representations less meaningful for the downstream tasks. Theoretically, we can have $2|\mathcal{T}|$ pretext tasks for $t \in \mathcal{T}$. Here, we believe two pretext tasks are sufficient to learn meaningful representations for the downstream tasks.

3.3 Integration with Instance Discrimination

As discussed in Sec. 2.1, contrastive learning can be viewed as a generalized pretext task in SSL. As contrastive learning has shown promising performance in

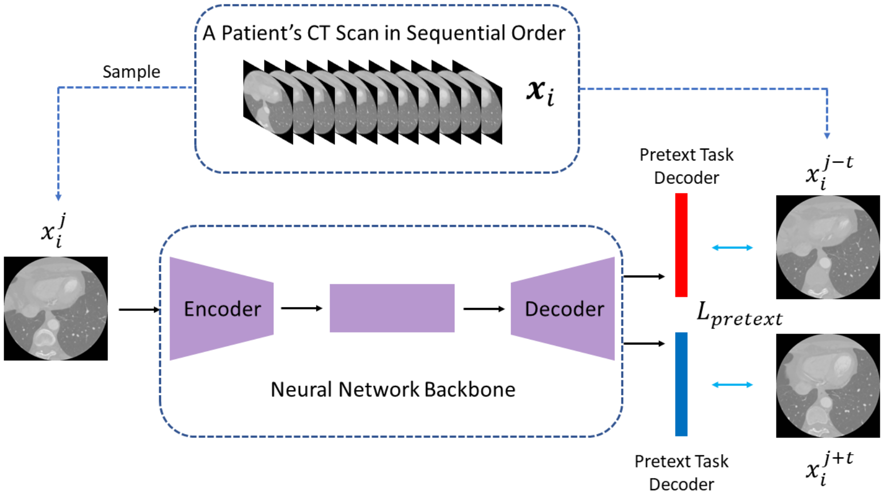


Fig. 5: Illustration of the proposed self-supervised multi-task learning framework. Given a sequence of medical images $\boldsymbol{x}_i$ (e.g. a CT scan of a patient), randomly sample x_i^j from the sequence and get corresponding x_i^{j-t} and x_i^{j+t} if possible. The architecture in purple is the neural network backbone that we are interested in (i.e. f_θ) and the two architectures in red and blue are two decoders for two pretext tasks respectively (i.e. g_{ϕ^-} and g_{ϕ^+}).

SSL for the encoder of image classification tasks, it is natural to consider including instance discrimination as the third pretext task in the MTL formulation when the neural network backbone can be perfectly decomposed as an encoder and a decoder. Here, we require that the encoder and the decoder can be trained independently (although we train them jointly).

For STOA contrastive learning methods, a positive pair is defined as two augmented views from the same instance and a negative pair is defined as two augmented views from two different instance. The common data augmentation policies include the combinations of cropping, resizing, flipping, color distortion by grayscale conversion, color distortion by jittering, cutout, Gaussian noise, Gaussian blurring, rotation, and Sobel filtering [6]. However, most of these data augmentation policies can not be applied to medical images directly for three reasons. First, medical images tend to be grayscale images or can only be transformed to grayscale images. Second, medical images are sensitive to local texture, which may be changed by the data augmentation. Third, unlike the random crops that contrastive learning methods usually work on, dense prediction tasks for medical images usually require the whole image as the input.

As discussed in Sec. 1, medical images share more similarities in the objects of interest than general objects in natural images due the human structural similarity. Instead of defining negative pairs, we only define positive pairs, inspired by Siamese networks [3, 8]. Based on the human structural similarity, we propose to define two slices from the same scan as a positive pair. Intuitively, we are using two *natural* variants of the same instance rather than *synthetic* variants (i.e. augmented views) of the same instance. For example, Fig. 2(a) and Fig. 2(d) can be viewed as a natural positive pair.

Formally, given an anchor image x_i^j, we define x_i^k as the variant of x_i^j. To avoid trivial solutions, we want x_i^j and x_i^k to be moderately different to increase the learning difficulty. We randomly sample k where we define $k \in \{j-2t, \cdots, j-t-1\} \cup \{j+t+1, \cdots, j+2t\}$. Let q denote the encoder which projects the input image to a feature vector. Here, the encoder could be a standard feature extractor used in image classification tasks, such as a ResNet feature extractor [20]. Following [8], two input images share the same encoder. Let h be a multi-layer perceptron (MLP), which further projects the encoded x_i^k to match the encoded x_i^j. Note, we use the stop gradient technique [19, 8] when encoding x_i^k. That is to say, we do not update the encoder when the loss backpropagates through $q(x_i^k)$, i.e. we use fixed weights for $q(x_i^k)$. By updating $q(x_i^j)$ alone, three pretext tasks can be optimized simultaneously. Given a positive pair x_i^j and x_i^k, we minimize the negative cosine similarity

$$\mathcal{L}_{sim} = -\frac{q(x_i^k)}{||q(x_i^k)||_2} \cdot \frac{h(q(x_i^j))}{||h(q(x_i^j))||_2} \tag{2}$$

, where each encoded feature vector is normalized by its l_2 norm. The motivation here is to learn the invariance between two images. Given two slices from the same patient, the invariance shared between two images is the general knowledge of human structures for the region of interest.

The final optimization object for the self-supervised MTL is to minimize the total loss of three pretext tasks. We have

$$\mathcal{L}_{self} = \mathcal{L}_{pretext} + \lambda \mathcal{L}_{sim} \tag{3}$$

, where λ is the hyperparameter to control the weight of $\mathcal{L}_{sim}$. In this work, to balance the weights among three pretext tasks, we use set $\lambda = 1$ [1]. The complete learning framework is present in Fig. 6. Note, for Sec. 3.2, there is no assumption of the architecture of the neural network, i.e. the learning framework should apply to any dense prediction task. However, for Sec. 3.3, we assume that the encoder should be the standard feature extractor for image classification tasks. We will empirically evaluate both frameworks in Sec. 4.

4 Experiments

We evaluate the proposed SSL frameworks on the whole heart segmentation (WHS) task. Unlike general semantic segmentation tasks that commonly take

[1] Task balancing is a topic of active research in MTL, which is beyond the scope of discussion in this work. We refer the interested readers to [37] for details.

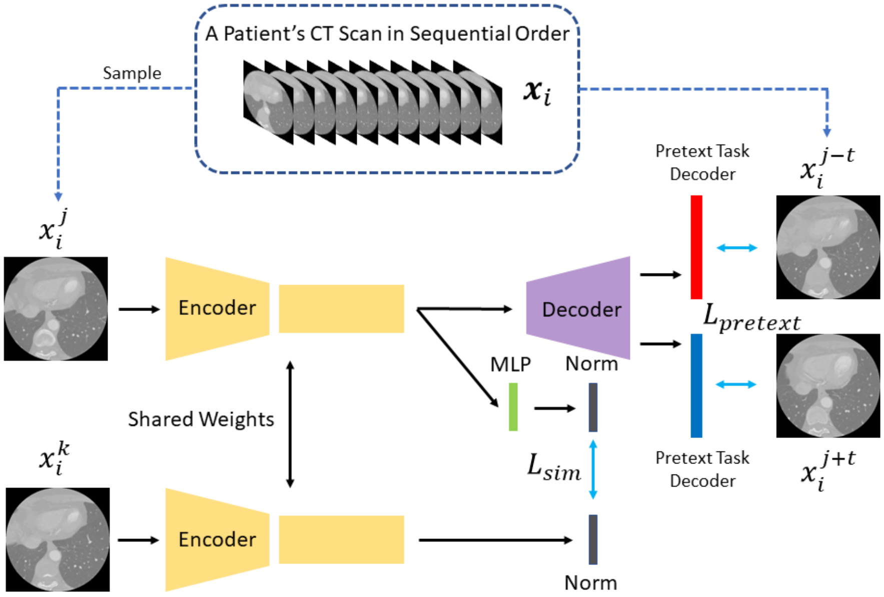


Fig. 6: Illustration of the proposed self-supervised multi-task learning framework with the additional instance discrimination pretext task. In addition to Fig. 5, we minimize the similarity between the encoded x_i^j and the encoded x_i^k, where two images share the same encoder. Note, we only update $q(x_i^j)$, the branch with the MLP in the MTL formulation. The neural network backbone that we are interested in includes an encoder (the architecture in yellow) and a decoder (the architecture in purple). The auxiliary MLP (the architecture in green) will not be used in the downstream tasks.

standard RGB images as input, WHS could have different source domains, namely CT scans and MRI scans. CT scans and MRI scans show variations in data modalities, which can be viewed in Fig. 3. The purposes of the experiments are twofold. First, we want to validate the theoretical advantages of the proposed framework. Second, we want to show that the proposed framework can extract meaningful representations on different type of data.

4.1 Datasets

We use two public benchmark datasets for WHS [2]. See Table 1 for the statistics of the datasets. Each dataset contains manual segmentation masks of 7 substructures of the heart for 20 patients: the left ventricle blood cavity, the right

[2] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/

Table 1: Dataset description.

dataset	axial	frontal	sagittal	# scans	# slices per scan	resolution
CT-WHS	✓	✗	✗	20	[177, 363]	512 × 512
MRI-WHS	✗	✗	✓	14	[120, 180]	[256, 340] × [256, 340]

ventricle blood cavity, the left atrium blood cavity, the right atrium blood cavity, the myocardium of the left ventricle, the ascending aorta, and the pulmonary artery. See Table 1 for the description of the datasets.

CT-WHS The CT-WHS dataset [42] is a benchmark dataset in WHS, which contains CT scans for 20 patients. CT-WHS has have axial views for all the patients. Each CT scan is represented as a 3D array and each slice of the scan is converted to a 2D image by mapping Hounsfield units to grayscale pixel values. CT-WHS The number of slices per patients differs across patients. Each slice has a fixed resolution 512 × 512.

MRI-WHS The MRI-WHS dataset [41] is a benchmark dataset in WHS, which contains MRI scans for 20 patients. Unlike CT-WHS, MRI-WHS have either frontal or sagittal views for each patient. Each scan is represented as a 3D array and each slice of the scan is converted into a 2D image by mapping MRI intensity values to grayscale pixel values. The image size and the number of slices per patients differs across patients.

4.2 Experimental Setup

Implementation For Sec. 3.2, we use a standard U-Net [32] as the neural network backbone. For simplicity, we replace the last convolutional layer of U-Net as two 1 convolutional layers with 64 channels for the input and 1 channel for the output. That is to say, we maximally shared the neural network backbone in MTL. For Sec. 3.3, as we need to decompose the neural network backbone into an encoder and a decoder, we choose FCN [24] as the neural network backbone. We use the ResNet50 [20] as the encoder. More precisely, the ResNet50 in this work denotes the ResNet50 architecture without the final fully-connected layer. The MLP consists of 2 fully-connected layers, whose number input channels and output channels are $2048 \mapsto 512$ and $512 \mapsto 2048$. The decoder is trained in the same way in Sec. 3.2. There is limited literature for utilizing human structural similarity in self-supervised multi-task learning for dense prediction tasks on medical images. For the baseline model, we use the SOTA contrastive learning framework MoCo V2 [7] [3]. For a fair comparison, we also use ResNet50 as the encoder. When fine-tuning with the downstream task, the decoder is randomly initialized. All models are implemented by PyTorch in a NVIDIA Tesla V100 GPU.

Hyperparameters For a fair comparison, we use the same set of hyperparameters for all models and all models are initialized with the same random seed.

[3] https://github.com/facebookresearch/moco

We use an Adam optimizer [21] with a fixed learning rate 10^{-3} across all experiments. The batch size is 16. Note, we do not use any data augmentation on the proposed framework. For MoCo V2, we use the default hyperparameters except K and the stochastic data augmentation policy. We set K as 1024 because we have much smaller datasets than ImageNet [4]. As discussed in Sec. 3.3, data augmentation plays an important role in SOTA contrastive learning frameworks. The original data augmentation policy is designed for RGB images instead of grayscale images. Here, we therefore use random flipping and stochastic Gaussian blurring as proposed in MoCo V2 [7].

Training and Evaluation For consistency among the datasets, all images and corresponding annotations are resized to 256×256. The resizing is also important for the baseline MoCo V2. The encoder for MoCo V2 is designed for standard ImageNet image size i.e.224 × 224. Each image is pre-processed by instance normalization. For the cardiac segmentation tasks on CT scans, we use all slices of 20 patients for the self-supervised pre-training. All models are pre-trained for the same of epochs as a fair comparison. In terms of the evaluation, the benchmark *linear classification protocol* [38, 6, 19, 26, 7] only applies to image classification tasks. Instead, we use the performance of supervised semantic segmentation as a *proxy* measurement for the quality of the learned representations. We choose four classes of interest in the cardiac segmentation task: the left ventricle blood cavity (LV), the right ventricle blood cavity (RV), the left atrium blood cavity (LA), and the right atrium blood cavity (RA). We split 20 patients into a training set of 5 CT scans and a test set of 15 CT scans. As in practical situations, the clinical annotators will only annotate a small amount of slices for each scan. We randomly sampled 20 annotated slices from each scan, with a total 100 slices as the training data. We use such as small training data to simulate the challenging data scarcity situation and also demonstrate the efficiency of the proposed framework. Given the self-supervised pre-trained neural network backbone, we fine-tune the model with the small training set and report the Intersection-Over-Union (IOU) of each class of interest and the mean IOU (mIOU) on the test set. The same training strategy applies for MRI scans. However, we only use 14 MRI scans with sagittal view as the self-supervised pre-training data. For the evaluation, we split 14 patients into a training set of 4 MRI scans and a test set of 10 MRIs. Similarly, we randomly sampled 25 annotated slices from each scan as the training images.

4.3 Results

We first evaluate the proposed framework in Sec. 3.2 on CT scans and MRI scans. We pre-train the U-Net on CT scans for 100 epochs and on MRI scans for 400 epochs as CT scans have around 3 times more slices than CT scans. As discussed in Sec. 4.2, we use the downstream task cardiac segmentation as a proxy evaluation. For CT scans, the models are trained with 100 labeled CT slices until convergence and tested with 4080 CT slices. For MRI scans, the

[4] K is originally 65532.

Table 2: Proxy evaluation of self-supervised multi-task learning on axial CT scans with U-Net as the neural network backbone.

Method	LV	RV	LA	RA	mIOU
w/o pre-training	0.445	0.378	0.427	0.429	0.420
SSMTL (t=5)	**0.656**	0.496	**0.657**	0.537	**0.586**
SSMTL (t=10)	0.571	0.478	0.545	**0.558**	0.537
SSMTL (t=15)	0.590	**0.499**	0.541	0.479	0.527

Table 3: Proxy evaluation of self-supervised multi-task learning on sagittal MRI scans with U-Net as the neural network backbone.

Method	LV	RV	LA	RA	mIOU
w/o pre-training	0.447	0.491	0.260	0.267	0.354
SSMTL (t=5)	0.552	0.489	**0.297**	**0.371**	**0.427**
SSMTL (t=10)	0.600	**0.509**	0.257	0.271	0.409
SSMTL (t=15)	**0.640**	**0.509**	0.283	0.234	0.417

number of training and test slices are 100 and 1480 respectively. The results of the segmentation performance are reported in Table 2 and Table 3. We denote the proposed framework as SSMTL. The U-Net pre-trained with SSMTL outperforms the U-Net without pre-training by a large margin on both datasets. In fact, this large margin is caused by the data scarcity. With insufficient labeled images, which is quite common in medical domain, traditional supervised approaches could easily fail. The proposed framework could be an efficient solution for this challenge. $t = 5$ gives the overall best performance in both tables, but in practice, the choice of t depends on the thickness of the slice. It is worth mentioning that the length of the sequence and the shape of the structure would influence the performance of representation learning. As shown in Table 2 and Table 3, the pre-training leads to more performance gain for scans with more sequential slices and larger structures. We also perform an ablation study for the number of pre-training epochs in Fig. 7. More pre-training epochs might not always help because the model could overfit the pretext tasks. This also leads to an interesting research question about how to measure the task affinity between pretext tasks and downstream tasks, which left as future work.

We repeat the previous experiment to evaluate the extended framework proposed in Sec. 3.3. We denote this extension of SSMTL as SSMTL+. We use $t = 5$. This time, we use a FCN with a ResNet50 encoder. The results are present in Table 4 and Table 5. Surprisingly, although a ResNet-FCN without pre-training shows much better results than its U-Net counterpart, U-Net pre-trained with SSMTL outperforms ResNet-FCN pre-trained with SSMTL+. Another interesting phenomenon is that ResNet-FCN pre-trained with SSMTL has a decreased performance. Both phenomena can be explained by the relationship between the network architecture and the target task. Note, there is a structural difference between U-Net and FCN, where U-Net has a balanced architecture between the

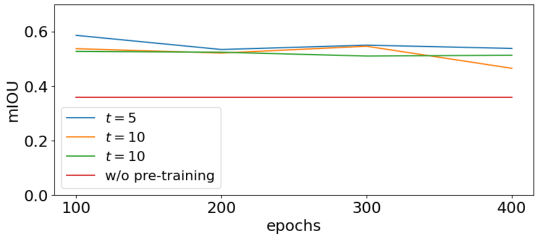


Fig. 7: The learning goal for pretext tasks might not be well-aligned with the learning goal for downstream tasks: more pre-training epochs do not always help the downstream tasks.

Table 4: Proxy evaluation of self-supervised multi-task learning on CT scans with ResNet-FCN as the neural network backbone.

Method	LV	RV	LA	RA	mIOU
w/o pre-training	0.548	0.475	0.491	0.421	0.483
MoCo V2	0.586	0.491	0.512	0.424	0.503
SSMTL	0.573	0.487	0.451	0.398	0.477
SSMTL+	0.607	0.487	0.523	0.434	0.513
SSMTL(U-Net)	**0.656**	**0.496**	**0.657**	**0.537**	**0.586**

encoder and the decoder but FCN puts more weight on the encoder. This enables ResNet-FCN to be more sensitive to semantic information, as ResNet50 is a seminal feature extractor, but also weakens its learning ability for dense prediction tasks with less semantic contents (i.e. no semantic labels). The pretext tasks proposed in Sec. 3.2 are not designed to extract semantic information as a segmentation task. So FCN might not be the suitable neural network backbone. SSMTL+ actually mitigates the issue with the additional instance discrimination task and shows slightly better performance than MoCo V2. Compared with benchmark SSL pre-training datasets such as ImageNet-1M [10] and Instagram-1B [25], the data scarcity in medical tasks will impair the performance of data-driven SSL models such as MoCo V2.

Denoising We have another ablation study to validate our hypothesis of the relationship between the architecture and the target task. Here, we examine the proposed framework with a simple downstream task, denoising. Denoising is a dense prediction task with pixel-to-pixel mapping. The pretext tasks are highly correlated with denoising as they are both reconstructing images. Moreover, there is no semantic labels involved. We use the same training and test split for CT scans. We add synthetic noise to the original images and use them as the training/test images. We treat the original images as the ground truth. Following [15], we implement the noise model as a zero-mean image-dependent Gaussian distribution. We utilize the models pre-trained from previous experiments. The training for the denoising downstream task is performed by minimizing the L1

Table 5: Proxy evaluation of self-supervised multi-task learning on sagittal MRI scans with ResNet-FCN as the neural network backbone.

Method	LV	RV	LA	RA	mIOU
w/o pre-training	0.445	0.415	0.186	0.254	0.325
MoCo V2	0.481	0.443	0.226	0.276	0.357
SSMTL	0.454	0.399	0.145	0.243	0.310
SSMTL+	0.501	0.473	0.211	0.257	0.361
SSMTL(U-Net)	**0.552**	**0.489**	**0.297**	**0.371**	**0.427**

Table 6: Proxy evaluation of self-supervised multi-task learning with denoising on CT scans.

Method	Network	PSNR
w/o pre-training	U-Net	36.04
SSMTL	U-Net	37.22
w/o pre-training	ResNet-FCN	34.85
MoCo V2	ResNet-FCN	32.85
SSMTL	ResNet-FCN	35.38
SSMTL+	ResNet-FCN	34.66

loss between the noisy input and the clean original images. We report the peak signal-to-noise ratio (PSNR) in Table 6, where U-Net outperforms FCN by a large margin.

Finally, we want to clarify that the experiments in this section are only used to validate the theoretical discussion in a simplified scenario. The proposed framework is designed for sequential medical images only. In practice, the medical tasks could have more complex problem settings and data challenges, which requires further consideration. In addition, we conclude that the choice of the neural network backbone should be dependent on the downstream dense prediction tasks. A possible future research direction could be using *neural architecture search* [13] to find the optimal network.

5 Conclusion

In this work, we propose a self-supervised representation learning framework for dense prediction tasks on sequential medical images. The proposed framework utilizes the human structural similarity to integrate MTL and SSL. The theoretical discussion and empirical analysis show that the proposed framework has several advantages over SOTA SSL methods, which are originally designed for natural images, on label-efficient medical image analysis. Limited by space, we only investigate a few downstream tasks on medical images. In the future, we will generalize the proposed framework for more dense prediction tasks in the medical domain and study the task affinity between the pretext tasks and the downstream tasks.

References

1. Baldi, P., Pineda, F.: Contrastive learning and neural oscillations. Neural Computation **3**(4), 526–545 (1991)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE TPAMI **35**(8), 1798–1828 (2013)
3. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. In: NeurIPS. pp. 737–744 (1993)
4. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**(4), 834–848 (2017)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
9. Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., Xing, E.P.: Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 263–273. Springer (2018)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE (2009)
11. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. pp. 1422–1430 (2015)
12. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: MICCAI. pp. 544–552 (2018)
13. Dong, N., Xu, M., Liang, X., Jiang, Y., Dai, W., Xing, E.: Neural architecture search for adversarial medical image segmentation. In: MICCAI. pp. 828–836 (2019)
14. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
15. Foi, A., Trimeche, M., Katkovnik, V., Egiazarian, K.: Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. IEEE TIP **17**(10), 1737–1754 (2008)
16. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
17. Gopnik, A., Meltzoff, A.N., Kuhl, P.K.: The scientist in the crib: Minds, brains, and how children learn. William Morrow & Co (1999)
18. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: AISTATS. pp. 297–304 (2010)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

22. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
25. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV. pp. 181–196 (2018)
26. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR. pp. 6707–6717 (2020)
27. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84. Springer (2016)
28. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
29. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: ECCV. pp. 762–780. Springer (2020)
30. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: ICML. pp. 759–766 (2007)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE TPAMI **39**(6), 1137–1149 (2016)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
33. de Sa, V.R.: Learning classification with unlabeled data. In: NeurIPS. pp. 112–119 (1994)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
35. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning. In: NeurIPS. vol. 33, pp. 6827–6839 (2020)
36. Vandenhende, S., Georgoulis, S., De Brabandere, B., Van Gool, L.: Branched multi-task networks: deciding what layers to share. In: British Machine Vision Conference (2020)
37. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE TPAMI (2021)
38. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR. pp. 3733–3742 (2018)
39. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. pp. 649–666. Springer (2016)
40. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: CVPR. pp. 4106–4115 (2019)
41. Zhuang, X., Bai, W., Song, J., Zhan, S., Qian, X., Shi, W., Lian, Y., Rueckert, D.: Multiatlas whole heart segmentation of ct data using conditional entropy for atlas ranking and selection. Medical Physics **42**(7), 3822–3833 (2015)
42. Zhuang, X., Rhode, K.S., Razavi, R.S., Hawkes, D.J., Ourselin, S.: A registration-based propagation framework for automatic whole heart segmentation of cardiac mri. IEEE TMI **29**(9), 1612–1625 (2010)