Joint Self-Supervised Image-Volume Representation Learning with Intra-Inter Contrastive Clustering

Duy M. H. Nguyen[1,10*], **Hoang Nguyen**[2], **Mai T. N. Truong**[3], **Tri Cao**[2], **Binh T. Nguyen**[2]
Nhat Ho[4], **Paul Swoboda**[5], **Shadi Albarqouni**[6,7], **Pengtao Xie**[8], **Daniel Sonntag**[9,10]

[1]University of Stuttgart, [2]AISIA Lab, University of Science -VNU HCM, [3]Dongguk University
[4]University of Texas at Austin [5]Max Planck Institute for Informatics, [6]Helmholtz AI, Helmholtz Munich, [7]University of Bonn,
[8]University of California San Diego, [9]Oldenburg University, [10]German Research Center for Artificial Intelligence
[*]Corresponding email: ho_minh_duy.nguyen@dfki.de

Abstract

Collecting large-scale medical datasets with fully annotated samples for training of deep networks is prohibitively expensive, especially for 3D volume data. Recent breakthroughs in self-supervised learning (SSL) offer the ability to overcome the lack of labeled training samples by learning feature representations from unlabeled data. However, most current SSL techniques in the medical field have been designed for either 2D images or 3D volumes. In practice, this restricts the capability to fully leverage unlabeled data from numerous sources, which may include both 2D and 3D data. Additionally, the use of these pre-trained networks is constrained to downstream tasks with compatible data dimensions. In this paper, we propose a novel framework for unsupervised joint learning on 2D and 3D data modalities. Given a set of 2D images or 2D slices extracted from 3D volumes, we construct an SSL task based on a 2D contrastive clustering problem for distinct classes. The 3D volumes are exploited by computing vectored embedding at each slice and then assembling a holistic feature through deformable self-attention mechanisms in Transformer, allowing incorporating long-range dependencies between slices inside 3D volumes. These holistic features are further utilized to define a novel 3D clustering agreement-based SSL task and masking embedding prediction inspired by pre-trained language models. Experiments on downstream tasks, such as 3D brain segmentation, lung nodule detection, 3D heart structures segmentation, and abnormal chest X-ray detection, demonstrate the effectiveness of our joint 2D and 3D SSL approach. We improve plain 2D Deep-ClusterV2 and SwAV by a significant margin and also surpass various modern 2D and 3D SSL approaches.

Introduction

Creating large-scale medical image datasets for training neural networks is a major obstacle due to the complexity of data acquisition, expensive annotations, and privacy concerns (Cheplygina, de Bruijne, and Pluim 2019; Kaissis et al. 2020). To alleviate these challenges, a conventional approach is to train deep networks, e.g., ResNet-50 (He et al. 2016), on large-scale natural image datasets such as ImageNet (Deng et al. 2009) and subsequently fine-tune them on the target medical domain. However, such schemes are suboptimal due to the large domain discrepancy between natural
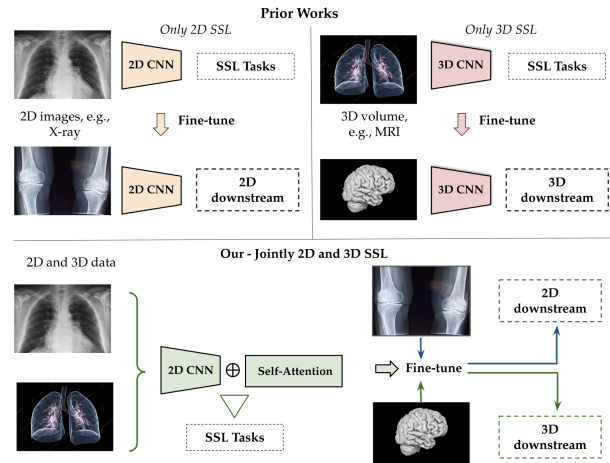
Figure 1: The main distinctions between our work and prior studies on 2D and 3D self-supervised learning. We can learn representations from diverse data and the pre-trained weights can be transferred for both 2D and 3D downstream tasks.

images and medical data (Raghu et al. 2019; Nguyen et al. 2022b). This has motivated other techniques for collecting annotated medical datasets across domains and training networks using full (Gibson et al. 2018; Chen, Ma, and Zheng 2019) or semi-supervision (Wang et al. 2020). Nevertheless, the amount of acquired relevant training data in this manner is still limited, which significantly limits the performance of deep neural networks.

Self-supervised learning (SSL) has recently emerged as a new trend in medical imaging due to its ability in obtaining feature representations from unlabeled data by solving proxy tasks, which can be broadly categorized into *generative* (Chen et al. 2019) and *discriminative* ones (Chen et al. 2020a; He et al. 2020). Discriminative SSL can be further separated into three directions: instance level-based methods (Zbontar et al. 2021; Caron et al. 2021), contrastive learning-based methods (He et al. 2020; Chen, Xie, and He 2021) and clustering-based methods (Caron et al. 2020; Li et al. 2021). Depending on a specific 2D, e.g., X-ray images or 3D magnetic resonance imaging (MRI) application, variations of these methods can be modified using 3D convolutional neural networks (CNNs) or Transformer architectures (Taleb et al. 2020; Haghighi et al. 2021; Tang et al. 2022).

However, all aforementioned SSL methods have been de-

signed to learn on either 2D or 3D data modalities. As a result, they suffer from two major limitations. First, the ability to exploit unlabeled data from multiple source domains, which commonly occurs in medical data, is restricted. For instance, 3D CNN-based SSL methods can not use X-rays, digital retinal, and dermoscopic images taken from lung, eye retina, and skin lesions, respectively. Although 2D CNN-based SSL methods can process 3D volumes slice-by-slice along a specific plane (either sagittal, coronal, or horizontal) (Nguyen et al. 2022a; Jun et al. 2021), these approaches do not capture long-range inter-slice correlations and thus may result in inferior performance in 3D applications. Second, using a purely 2D or 3D strategy limits the fine-tuning phase since the pre-trained models are only applicable for downstream tasks with the same dimensionality. For instance, pre-trained 3D-CNN cannot handle object detection (Nguyen et al. 2021a, 2022c) (Table 4, third column) and similarly pre-trained 2D-CNN might not be usable for 3D classification tasks (Table 3, second column).

In this work, we propose a novel technique to overcome those barriers by presenting a hybrid SSL architecture harnessing both 2D and 3D medical data. The method has the following properties. First, it is built on top of cutting-edge 2D SSL baselines while reserving designed CNN architecture, benefiting from the latest advancements of SSL in natural images. Second, when applied to 3D data, we formulate both intra-dependencies inside slices and long-range inter-dependencies across slices, resulting in more complex contrastive cues that force the network to seek associated local and global feature representations.

Specifically, we compose a joint image-volume representation learning comprising a 2D CNN (ResNet-50) to extract feature embedding at the image level and a deformable attention transformer (Zhu et al. 2020; Liu et al. 2021; Xia et al. 2022) to express correlations among local slices, aiming to derive a holistic representation at the 3D volume level. Unlike standard attentions in Transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020) which treat all attention positions equally, our deformable mechanism pays attention to only a flexible small set of major slices conditioned on input data. This largely reduces computational complexity and permits handling the multi-scale feature maps which are desired properties in medical downstream tasks.

The proposed method is trained on SSL tasks utilizing both current 2D SSL methodologies and our two novel 3D pre-text tasks. To this end, we employ two state-of-the-art contrastive clustering-based SSL approaches, Deep-Cluster-V2 (Caron et al. 2018) and SwAV (Caron et al. 2020). With each baseline, we first perform the relevant 2D proxy tasks based on an *agreement clustering for 2D slices* taken from 3D volumes. We next compute multi-level features at each slice within a 3D volume encoded with their positions and feed them into the deformable transformer. The global embedded features derived from this transformer are employed to define an *agreement clustering for 3D volumes* and a *masked encoding feature prediction* motivated by the success of the language model BERT (Devlin et al. 2018). By optimizing these conditions, intuitively we are able to learn feature extractors at the local- and global-level

in a constraint manner, resulting in consistent cues and improved performance in downstream tasks. Furthermore, the pre-trained networks are adaptable with data dimensional compatibility by employing the 2D CNN for 2D tasks or the hybrid 2D CNN- Transformer architectures for 3D tasks.

In summary, we make the following contributions. First, we present an SSL framework capable of using various data dimensions and producing versatile pre-trained weights for both 2D and 3D downstream applications (Figure 1). Second, we introduce the deformable self-attention mechanisms which utilize multi-level feature maps and capture flexible correlations between 2D slices, resulting in a powerful global feature representation. On top of this, we developed the novel 3D agreement clustering extended from the earlier 2D clustering problem as well as proposed the masking embedding prediction. Finally, extensive experiments on public benchmarks confirmed that we improve state-of-the-art 2D baselines and surpass several latest SSL competitors based on CNN or Transformer.

Related Work

Self-supervised Learning in Medical Image Analysis
Our work is closely related to instance-based constrative learning and unsupervised contrastive clustering. The *instance-based contrastive methods* seek an embedding space where transformed samples, e.g., crops, drawn from the same instance, e.g., image, are pulled closer, and samples from distinct instances are pushed far away. The contrastive loss is constructed based on positive and negative feature pairs generated by various approaches, such as memory bank (Wu et al. 2018), end-to-end (Chen et al. 2020a), or momentum encoder (Chen, Xie, and He 2021). Despite achieving good performance in various settings, the instance-based method has crucial limitations in requiring a large negative batch size and choosing hard enough negative ones. The *unsupervised contrastive clustering* (Caron et al. 2018, 2020) in other directions tries to learn representations based on groups of images with similar features rather than individual instances. For instance, SwAV (Caron et al. 2020) simultaneously clusters the data while imposing consistency between cluster assignments generated for distinct augmentations of the same image. Currently, extensions on this direction have considered latent variables of centre points (Li et al. 2021), multi-view clustering (Pan and Kang 2021), or mutual information (Do, Tran, and Venkatesh 2021).

In medical image analysis, several SSL methods have designed pre-text tasks based on 3D volume's properties such as reconstructing spatial context (Zhuang et al. 2019), random permutation prediction (Chen et al. 2019), self-discovery and self-restoration (Zhou et al. 2021b; Haghighi et al. 2021). Some other efforts attempted to develop 3D CNN architecture while retaining defined SSL tasks on 2D CNN (Taleb et al. 2020). Another line of research considered the cross-domain training with two or more datasets, aiming to derive a generic invariant pre-trained model (Zhang et al. 2020). Besides, existing methods also exploit the domain- and problem-specific cues such as structural similarity across 3D volumes in order to define global and local contrastive losses (Chaitanya et al. 2020; Xie et al. 2020).

However, most of these techniques have only been applied to 2D or 3D data, which are different from ours in terms of data usage and flexible pre-trained weights in downstream tasks (Figure 1).

SSL Transformer in Medical Imaging Vision transformers, adapted from sequence-to-sequence modeling in natural language processing, are initially used in image classification tasks (Dosovitskiy et al. 2020). In the context of SSL, 2D transformer-based methods such as Moco-v3 (Chen, Xie, and He 2021) and DINO (Caron et al. 2021) are also introduced and achieved promising performance. To elaborate 3D volumes, Tang et al. (2022) introduced a 3D transformer-based model comprising a Swin Transformer encoder (Liu et al. 2021) and skip connections. Likewise, Xie et al. (2021) adapted a mixed 2D-3D Pyramid Vision Transformer architecture (Wang et al. 2021) to learn rich representations from diverse data.

Compared with prior works in SSL (Caron et al. 2021; Tang et al. 2022), we employ Transformer to define the interaction between 2D slices inside a 3D volume rather than a fixed 2D or 3D network backbone, allowing us to adapt to varied data dimension downstream applications. Furthermore, we the first adapt deformable attention mechanism (Zhu et al. 2020; Liu et al. 2021; Xia et al. 2022) for SSL, which currently are only validated performance in supervised learning. Xie et al. (2021) shares the same ideas with us in jointly learning diverse unlabeled data; however, this method designs a specific SSL task while our 3D loss is extended directly from standard 2D cases. Also, we achieve similar or better performance compared with this baseline while using a smaller amount of unlabeled data.

Methodology

Our approach is built on top of 2D contrastive clustering learning baselines including Deep-ClusterV2 (Caron et al. 2018) and SwAV (Caron et al. 2020). Both approaches rely on clustering together features produced by neural network backbones. Deep-ClusterV2 forces each cluster to have roughly the same size. SwAV additionally imposes losses on assigning augmentations of an image into the same cluster. Below, we recapitulate the SwAV baseline and then show how it can be extended through the deformable self-attention (Zhu et al. 2020; Xia et al. 2022) to 3D volumes. Additionally, we introduce a new proxy task based on missing embedding prediction in order to make the designed architecture be stable under perturbations. An illustration of our approach can be seen in Figure 2. A variation of our method using DeepCluster-V2 can be derived analogously.

Notation: We assume to be given K unlabeled datasets $\mathbb{D} = \{D_1, D_2, ..., D_K\}$ consisting of instances $D_i = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_{m_i}}\}, i \in [1, K]$, which include $\mathbf{m_i}$ 2D or 3D volumes $\mathbf{X_j}, \mathbf{j} \in [\mathbf{1}, \mathbf{m_i}]$. Given a particular dataset $D \in \mathbb{D}$, we assume that each 3D volume contains n slices, i.e. $\forall \mathbf{X} \in D, \mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$.

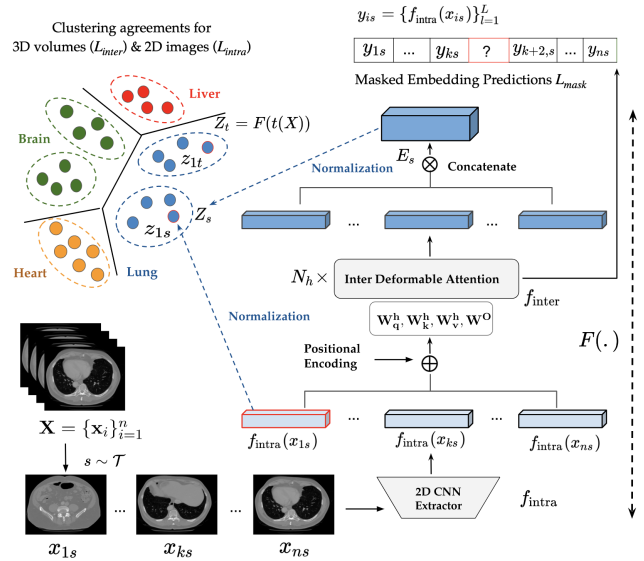Figure 2: Overview of our joint SSL image-volume framework. Given a 3D volume $\mathbf{X}$ and a random transformation s, we compute the embedding feature for each slice using a 2D-CNN extractor f_{intra} and produce a global feature Z_s through the Inter Deformable Attention f_{inter}. Similarly, corresponding features can be derived from 2D and 3D augmented views of $\mathbf{X}$ by another transformation t. Through cluster agreement losses for 2D slices (L_{intra}), e.g. between z_{1s} and z_{1t}, and for 3D volumes between Z_s and Z_t (L_{inter}), feature representations can be learned. Additionally, we employ a masked feature embedding prediction given 2D slices' embedding outputs as an SSL task to capture data's long-term interdependence.

Clustering Agreement for 2D Images

SwAV uses a proxy task for a "swapped" prediction problem in which the cluster assignment of a transformed image is to be found from the feature representation of another transformation of the same image and vice versa. In our framework, we refer to this proxy task as an *intra-dependence* correlation since it learns only from 2D slices inside a 3D volume without taking into account correlations between different slices of the same volume. Below we formally specify the intra-dependence correlation.

Let f_{intra} be a CNN, e.g., ResNet-50 (He et al. 2016), extracting feature embeddings for each 2D slice $\mathbf{x}_i \in \mathbf{X}$. The cluster assignment matrix $\mathbf{C} = [c_1, ..., c_H]$ has columns c_j, each column corresponding to the feature representation of the j-th cluster, and H is the number of hidden clusters. Given a 2D slice $\mathbf{x}_i \in \mathbf{X}$, we choose randomly two transformations $s, t \in T$, where T is a set of pre-defined image transformations. We apply s and t on $\mathbf{x}_i$ and obtain two augmented views $\mathbf{x}_{is}, \mathbf{x}_{it}$. Using f_{intra} and normalization gives us the respective features $\mathbf{z}_{it}$ and $\mathbf{z}_{is}$ (Figure 2), i.e.

$$\mathbf{z_{ik}} = f_{\text{intra}}(\mathbf{x}_{ik})/||f_{\text{intra}}(\mathbf{x}_{ik})||_2, \ k \in \{s, t\}. \quad (1)$$

These features are then used to find corresponding cluster assignments $\mathbf{q}_{it}, \mathbf{q}_{is}$, i.e., the probability distribution over all clusters, called codes in SwAV. To find these codes, we sample a batch of size B from slices of volumes coming from all datasets and optimize

$$\max_{\mathbf{Q} \in \mathbb{Q}} \mathbf{Tr}(\mathbf{Q}^T \mathbf{C}^T \mathbf{Z}) + \epsilon H(\mathbf{Q}), \quad (2)$$

where $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_{2B}]$ is formed by adding features z_{it}, z_{is} of each x_i in the batch B, the assignment matrix is $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{2B}]$ and $\mathbb{Q} = \{\mathbf{Q} \in \mathbb{R}_+^{K \times B} : \mathbf{Q}\mathbb{1}_B = \frac{1}{K}\mathbb{1}_K, \mathbf{Q}\mathbb{1}_K = \frac{1}{B}\mathbb{1}_B\}$ is the set of all possible assignment matrices such that slices are assigned on average uniformly, H is the entropy function and ϵ is a hyper-parameter that controls the smoothness of the mapping. Since views coming from the same sample $\mathbf{x}_i$ should have features that are assigned to the same cluster, we formulate the intra-dependency code prediction loss

$$L_{\text{intra}}(\mathbf{z}_{it}, \mathbf{q}_{it}, \mathbf{z}_{is}, \mathbf{q}_{is}) = l(\mathbf{z}_{it}, \mathbf{q}_{is}) + l(\mathbf{z}_{is}, \mathbf{q}_{it}) \quad (3)$$

where the function $l(\mathbf{z}, \mathbf{q})$ quantifies the fit between feature $\mathbf{z}$ and code assignment $\mathbf{q}$ defined as

$$l(\mathbf{z}_t, \mathbf{q}_s) = -\sum_k \mathbf{q}_s^k \log \mathbf{p}_t^k, \text{ where } \mathbf{p}_t^k = \frac{\exp(\frac{1}{\tau}\mathbf{z}_t^T \mathbf{c}_k)}{\sum_{k'} \exp(\frac{1}{\tau}\mathbf{z}_t^T \mathbf{c}_{k'})}.$$
$$(4)$$

Here τ is a hyper-parameter.

Intuitively, if two features encode views coming from the same slice, the loss $l(\mathbf{z}_t, \mathbf{q}_s)$ in Eq. (4) encourages their predicted clusters should be identical. Finally, by optimizing Eq. (3) over $\mathbf{x}_i \in \mathbf{X}$ we can learn feature representations f_{intra} and centroids $\mathbf{C}$ by minimizing

$$L_{2D} = \min_{f_{\text{intra}}, \mathbf{C}} \mathbb{E}_{\mathbf{x}_i \in \mathbf{X}} \left[L_{\text{intra}}(\mathbf{z}_{it}, \mathbf{q}_{it}, \mathbf{z}_{is}, \mathbf{q}_{is}) \right], s, t \sim T.$$
$$(5)$$

Clustering Agreement for 3D Volumes with Inter Deformable Attention

In the presence of both unlabeled 2D and 3D data, we argue that the clustering agreement constraint in Eq.(4) should also hold for feature representations of *different views of the 3D volume* (Figure 2). We call this agreement as an *inter-dependence correlation*. It forces the feature representation to additionally consider long-range interactions among 2D slices inside a 3D volume (Eq.(10)). To this end, we adapt the Transformer to aggregate local features computed by f_{intra} at each slice to form a holistic feature representation for a 3D volume. However the standard attention mechanisms in vanilla Transformer such as ViT (Dosovitskiy et al. 2020) does not fit well in our setting when it permits excessive number of keys to contribute per query patch. As a result, the required memory and computational costs increase significantly as well as features can be influenced by irrelevant parts.

To mitigate these problems, we use the deformable self-attention mechanism which is recently introduced in supervised learning such as object detection and image classification (Zhu et al. 2020; Xia et al. 2022). Generally this strategy seeks important positions of keys and value pairs in self-attention in a dependent-way rather than a fixed window size as ViT (Figure 3). Specifically, these important regions are learnt using an offset network that takes input query features and returns corresponding offsets whose regions subsequently are used to sample candidates keys/values (Figure 3). In this work, we use this deformable attention to SSL for the first time, aiming to learn the association among feature

embedding of 2D slices. We call this as Inter Deformable Attention and denote by f_{inter}. The f_{inter} contains N iden-
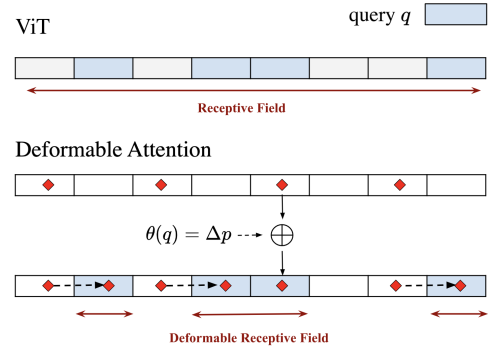

Figure 3: Comparison of Deformable Attention (DAT) with standard Vision Transformer (ViT) in our setting using slice's embedding vector. Given a query q, ViT pays attention to all possible positions including possibly less relevant feature maps while DAT learns important regions based on grid points (red points) and their shifted vectors using offsets Δp predicted by $\theta(q)$.

tical stacked layers. Each layer is composed of multi-head attention (MHA) layer followed by a simple feed-forward layer. Given an input tensor $Y \in \mathbb{R}^{D \times F_{\text{in}}}$ added with a positional encoding to provide order information, the output of a single head h at each layer using deformable attention can be computed by the following step.

$$q^{(h)} = YW_q^h, \ \tilde{k}^{(h)} = \tilde{Y}W_k^h, \ \tilde{v}^{(h)} = \tilde{Y}W_v^h \quad (6)$$

$$\text{with } \tilde{Y} = \phi(Y; p + \Delta p), \ \Delta p = \theta_{\text{offset}}\left(q^{(h)}\right) \quad (7)$$

where W_q^h, W_k^h and $W_v^h \in \mathbb{R}^{F_{\text{in}} \times d_h^v}$ are learned linear transformation that map the input Y to queries, keys, values respectively; θ_{offset} be the offset network that takes input as queries q^h and returns the offsets Δp; $p \in \mathbb{R}^{D_G \times 2}$ denotes for the uniform grid of points with $D_G = D/r$ by a factor r to down-sample the grid size; finally $\phi(.;.)$ be a differentiable linear interpolation function used to sample important key/queries pairs inside predicted offsets.

We now compute the output of a deformable attention head h as:

$$O^{(h)} = \sigma\left(q^{(h)}\tilde{k}^{(h)\top}/\sqrt{d^{(h)}} + \phi(\hat{B}; R)\right)\tilde{v}^{(h)} \quad (8)$$

where $\sigma(.)$ denotes the softmax function, $d^{(h)}$ is the dimension of each head h, $\hat{B} \in \mathbb{R}^{(2D-1)}$ be a relative position bias matrix, R be the relative position offsets. More details on this bias matrix, we refer the readers to (Liu et al. 2021; Xia et al. 2022). The outputs of all heads (MHA) are aggregated by concatenating and projecting again as MHA = Concat $\left[O^{(1)}, ..., O^{(Nh)}\right]W^O$ where $W^O \in \mathbb{R}^{d_v \times d_v}$ is another learned linear transformation and Nh is the number of heads.

Given defined f_{inter}, we construct a 3D feature representation $\mathbf{Z}_s$ for an augmented view $\mathbf{X}_s = \{\mathbf{x}_{1s}, \mathbf{x}_{2s}, ..., \mathbf{x}_{ns}\}$ of $\mathbf{X}$ as follows. We denote by

$$\mathbf{Y} = \left[\{f_{\text{intra}}(\mathbf{x}_{1s})\}_{l=1}^L, \ldots, \{f_{\text{intra}}(\mathbf{x}_{ns})\}_{l=1}^L\right] \quad (9)$$

be the stacked input vectors with $\{f_{\text{intra}}(\mathbf{x}_{is})\}_{l=1}^L$, $i \in [1, n]$ indicates the multi-level features of image $\mathbf{x}_{is}$ taken from the L last layers in f_{intra}. We then normalize the ouput of f_{inter} and obtain

$$\mathbf{Z_s} = f_{\text{inter}}(\mathbf{Y})/||f_{\text{inter}}(\mathbf{Y})||_2 \quad (10)$$

which is the holistic feature of $\mathbf{X}_s$. The embedding $\mathbf{Z}_t$ for transformation $t \in T$ is computed analoguously. The clustering agreement for 3D volumes generalized from Eq.(3) can be defined as

$$L_{\text{inter}}(\mathbf{Z}_t, \mathbf{q}_t^{3D}, \mathbf{Z}_s, \mathbf{q}_s^{3D}) = l(\mathbf{Z}_t, \mathbf{q}_s^{3D}) + l(\mathbf{Z}_s, \mathbf{q}_t^{3D}) \quad (11)$$

where $\mathbf{q}_s^{3D}, \mathbf{q}_t^{3D}$ are codes of $\mathbf{Z}_s, \mathbf{Z}_t$ obtained by solving the matching problem in Eq.(2) where inputs are 3D augmented views' feature represents across 3D volumes $\mathbf{X_i}$ in a batch size $B \in \mathbb{D}$. Intuitively, two 3D features $\mathbf{Z}_s$ and $\mathbf{Z}_t$ should be identical in their cluster assignments. Finally, by minimizing over samples in $\mathbb{D}$, we jointly learn both $f_{\text{intra}}, f_{\text{inter}}$ and $\mathbf{C}$ through

$$L_{3D} = \min_{f_{\text{intra}},\mathbf{C},f_{\text{inter}}} \mathbf{E}_{\mathbf{X} \in \mathbb{D}} \left[L_{\text{inter}}(\mathbf{Z}_t, \mathbf{q}_t^{3D}, \mathbf{Z}_s, \mathbf{q}_s^{3D}) \right]$$
$$\text{with } s, t \sim T. \quad (12)$$

Masked Feature Embedding Prediction

To enhance long-term dependence learning of f_{inter}, we additionally introduce a new SSL proxy task inspired by the BERT language model (Devlin et al. 2018). Given a set of 2D slice embedding vectors $\mathbf{Y}$ in Eq.(9) obtained from $\mathbf{X_s}$ ($\mathbf{X} \in \mathbb{D}, s \sim T$), we dynamically mask some inputs $\{f_{\text{intra}}(\mathbf{x}_{is})\}_{l=1}^L$, $i \in [1, n]$ and ask the *Inter Deformable Attention* to predict missing encoding vectors given the unmasked embedding vectors. To do this, we define a binary vector $\mathbf{m} = (m_1, \ldots, m_n)$ of length n where $m_i = 1$ indicate the input i-th of $\mathbf{Y}$ will be masked and 0 otherwise. The input for SSL task then is defined as:

$$\mathbf{m} \odot \mathbf{Y} = \begin{cases} [\text{MASK}], & m_i = 1 \\ \{f_{\text{intra}}(\mathbf{x}_{is})\}_{l=1}^L, & m_i = 0 \end{cases} \quad (13)$$

where MASK is a learnable parameter during the training step. We denote by f_{decode}, a fully connected layer, that takes the outputs of f_{inter} and predicts masked vectors. For each $\mathbf{m}$, we randomly assign $m_i = 1$ for 10% of $\mathbf{m}$. The output of f_{decode} at each masked $\mathbf{y}_i$ is:

$$\mathbf{y}_i = \mathbf{W_d}\mathbf{h}_i^N + \mathbf{b}_i, \quad \text{where } m_i = 1. \quad (14)$$

with $\mathbf{W}_d \in \mathbb{R}^{F_{in} \times F_D}$ and $\mathbf{b}_i \in \mathbb{R}^{F_{in}}$ are fully-connected layers and biases respectively. The masked feature embedding prediction is defined as:

$$L_{mask} = \min_{\substack{f_{\text{inter}} \\ f_{\text{decode}}}} \mathbf{E}_{\substack{\mathbf{x} \in \mathbb{D} \\ s \sim T}} \left[\sum_{i:m_i=1} ||f_{\text{intra}}(\mathbf{x}_{is}) \right.$$
$$\left. - f_{\text{decode}}(f_{\text{inter}}(\mathbf{m} \odot \mathbf{Y}))||_2 \right] \quad (15)$$

Experiment Results

Data and Baseline Setup

Pre-training and Downstream Tasks We describe the details of datasets used for pre-training and downstream

tasks in Table 1 and Table 2, respectively. In summary, there are thirteen datasets comprising LUNA2016 (Setio et al. 2015), LiTS2017 (Bilic et al. 2019), BraTS2018 (Bakas et al. 2018), MSD (Heart) (Simpson et al. 2019), MOTS (Zhang et al. 2021), LIDC-IDRI (Clark et al. 2013; Armato III et al. 2011), RibFrac (Jin et al. 2020), TCIA-CT (Clark et al. 2013; Harmon et al. 2020), NIH ChestX-ray8 (Wang et al. 2017), MMWHS-CT/MMWHS-MRI (Zhuang and Shen 2016), VinDR-CXR (Nguyen et al. 2022c), and JSRT (Shiraishi et al. 2000; Van Ginneken, Stegmann, and Loog 2006). In pre-training settings, we mainly evaluate in two scenarios, namely *Universal* and *Unified* following prior works of Zhang et al. (2020) and Xie et al. (2021), respectively. However, we cannot access the dataset called "Tianchi dataset" in *Unified* setting thus we only train with five remaining datasets. The downstream tasks are conducted in three contexts with diverse applications as described Table 2. For objective assessment, we use Intersection over Union (IoU) computed on 3D data for segmentation, Area Under the Curve (AUC) for 3D classification, Dice coefficient scores for 2D segmentation, and Average Precision with IoU=0.5 for multi-object detection.

Table 1: Overview pre-training settings in our experiment. The *Universal* setting uses four unlabeled 3D datasets while *Unified* uses six unlabeled datasets including mixed 2D and 3D modalities.

Setting	Pre-Training Data	Modality	Num	Access
Universal	LUNA2016	3D CT	623	✓
	LiTS2017	3D CT	111	✓
	BraTS2018	3D MRI	760	✓
	MSD (Heart)	3D MRI	30	✓
Unified	MOTS	3D CT	936	✓
	LIDC-IDRI	3D CT	1008	✓
	Tianchi	3D CT	N/A	✗
	RibFrac	3D CT	420	✓
	TCIA-CT	3D CT	1300	✓
	NIH ChestX-ray8	2D Xrays	108948	✓

Table 2: Overview downstream tasks used in our experiment. **Seen Domain** indicates for downstream tasks where the training data was used in the pre-training step without labels, **Unseen Domain** means that datasets in pre-training and downstream are different.

Setting	Testing Data	Modality	Num	Pre-training	Task
Seen domain in Universal	BraTS2018	3D MRI	285	Universal	Tumor Segmentation
	LUNA 2016	3D CT	888	Universal	Lung Nodes Classification
Unseen Domain in Universal	MMWHS-CT	3D CT	20	Universal	Heart Structures Segmentation
	MMWHS-MRI	3D MRI	20	Universal	Heart Structures Segmentation
	VinDR-CXR	2D X-ray	4394	Universal	Abnormal Chest Detetction
Unseen Domain in Unified	JSRT	2D X-ray	247	Unified	Multi-Organ Segmentation

Competing Algorithms We implement variations of Deepcluster and SwAV based the proposed method and compare with the following approaches:

- *2D SSL methods*: SimCLR (Chen et al. 2020a), PGL (Xie et al. 2020), Moco-v2 (Chen et al. 2020b), Deep-Cluster-v2 (Caron et al. 2020), SwAV (Caron et al. 2020), Barlow-Twins (Zbontar et al. 2021), Moco-V3 (Chen, Xie, and He 2021), PCRL (Zhou et al. 2021a), and DINO (Caron et al. 2021). Both Moco-v3 and DINO use Pyramid Transformer Unet (Xie et al. 2021) as backbone.
- *3D SSL methods*: 3D Rotation, 3D JigSaw (Taleb et al. 2020), Universal Model (Zhang et al. 2020), Models

Genesis (Zhou et al. 2021b), TransVW (Haghighi et al. 2021), SwinViT3D (Tang et al. 2022), and our two implementations for the 3D case of Deepcluster-v2 and SwAV, namely 3D-Deepcluster and 3D-SwAV.

- *2D/3D supervised pre-trained methods*: 2D pre-trained ImageNet (He et al. 2016), I3D (Carreira and Zisserman 2017), NiftyNet (Gibson et al. 2018), and Med3D (Chen, Ma, and Zheng 2019).
- *Other methods*: training from scratch for 2D or 3D using ResNet-50, V-Net architecture (Milletari, Navab, and Ahmadi 2016), 3D-Transformer (Hatamizadeh et al. 2022), Pyramid Transformer Unet (PTU) (Xie et al. 2021) and finally USST (Xie et al. 2021), a joint 2D and 3D approach similar to ours.

Most baseline results are taken from (Zhang et al. 2020) and (Xie et al. 2021). With LUNA2016 dataset, we use the latest ground-truth, denoted as LUNA2016-v2, and provide results obtained when training with batch sizes of 8, 16, 32, each with two trial times. For new competitors, we describe experiment setups in the appendix. In short, for 2D self-supervised methods (ResNet-50 backbone) such as Moco-v2 or Barlow-Twins, we extract all 2D slices from 3D volumes in pre-training data and train SSL tasks with 100 epochs. With state-of-the-art 3D SSL methods TransVW and Swin-ViT3D, we download pre-trained weights and use published implementation to fine-tune as author's suggestions. For two our implementations of 3D-Deepcluster and 3D-SwAV, we train with all 3D data of *Universal* in pre-training step.

Table 3: Comparing SSL approaches on **Seen Domains** trained on the *Universal setting*. Two top results in 2D or combined 2D-3D SSL data are red, blue. The best values in 3D-based methods and overall are in bold and underlined respectively. N/A indicates pre-trained models that are unable to transfer (Universal Model's results are not available in LUNA2016-v2).

Pre-training	Method	BraTS2018 (MRI - Segmentation)	LUNA2016-v2 (CT - Classification)
N/A	Scratch (3D)	58.51 ± 2.61	94.15 ± 3.97
	V-Net	59.01 ± 2.59	95.85 ± 1.09
	3D-Transformer	66.54 ± 0.40	85.15 ± 2.62
3D Supervised	I3D	67.83 ± 0.75	92.43 ± 2.63
	NiftyNet	60.78 ± 1.60	94.16 ± 1.52
	Med3D	66.09 ± 1.35	91.32 ± 1.47
3D Self-supervised	3D-Rotation	56.48 ± 1.78	**95.91** ± **1.26**
	3D-JigSaw	59.65 ± 0.81	89.12 ± 1.71
	Models Genesis	67.96 ± 1.29	92.46 ± 5.54
	Universal Model	**72.10** ± **0.67**	N/A
	3D-DeepCluster	59.20 ± 1.69	89.03 ± 2.56
	3D-SwAV	62.81 ± 1.03	88.79 ± 5.48
	TransVW	68.82 ± 0.38	93.84 ± 6.73
	SwinViT3D	70.58 ± 1.27	88.68 ± 2.63
N/A	Scratch (2D)	66.82 ± 1.32	N/A
2D Supervised	Pre-trained ImageNet	71.24 ± 2.30	N/A
2D Self-supervised	SimCLR	70.37 ± 1.11	N/A
	Moco-v2	70.82 ± 0.22	N/A
	Barlow-Twins	67.35 ± 0.55	N/A
	Deep-Cluster-v2	69.21 ± 2.10	N/A
	SwAV	69.83 ± 2.44	N/A
2D & 3D Self-supervised	**Our (Deep-Cluster-v2)**	72.81 ± 0.15	93.91 ± 0.67
	Our (SwAV)	73.03 ± 0.42	94.22 ± 1.11

Implementation Details

Pre-training Our method is trained in three stages. Stage 1 learns f_{intra} using Eq. (5) with 100 epochs using batch size of 1024 images, Stage 2 learns f_{inter} using Eq. (15) with 100 epochs using batch size of 12 volumes, and Stage 3 learns for both f_{intra}, f_{inter} using Eq. (12) also with 100 epochs and batch size of 12 volumes.

Table 4: Comparing SSL approaches on **Unseen Domains** trained on the *Universal setting*. Two top results in 2D or combined 2D-3D SSL data are red, blue. The best values in 3D-based methods and overall are in bold and underlined respectively. N/A indicates for pre-trained models that are unable to transfer.

Pre-training	Method	MMWHS (CT - Segm.)	MMWHS (MRI - Segm.)	VinDr-CXR (X-ray - Detect.)
N/A	Scratch (3D)	68.29 ± 1.68	67.04 ± 2.18	N/A
	V-Net	69.66 ± 3.65	67.50 ± 3.76	N/A
	3D-Transformer	67.30 ± 2.29	67.64 ± 2.21	N/A
3D Supervised	I3D	76.63 ± 2.32	66.71 ± 1.27	N/A
	Nifty Net	74.91 ± 2.78	64.60 ± 1.96	N/A
	Med3D	75.01 ± 0.74	63.43 ± 0.61	N/A
3D Self-supervised	3D Rotation	67.54 ± 2.80	71.36 ± 1.70	N/A
	3D Jigsaw	68.40 ± 2.92	72.99 ± 2.54	N/A
	Model Geneis	76.48 ± 2.89	74.53 ± 1.69	N/A
	Universal Model	78.14 ± 0.77	77.52 ± 0.50	N/A
	3D-DeepCluster	69.47 ± 1.44	75.83 ± 2.29	N/A
	3D-SwAV	69.90 ± 1.31	69.41 ± 1.93	N/A
	TransVW	**79.74** ± **2.78**	75.08 ± 2.04	N/A
	SwinViT3D	70.19 ± 1.23	**78.25** ± **1.66**	N/A
N/A	Scratch (2D)	74.25 ± 2.05	52.34 ± 4.31	24.35 ± 0.04
2D Supervised	Pre-trained ImageNet	73.49 ± 3.15	72.66 ± 2.46	27.82 ± 0.29
2D Self-supervised	SimCLR	78.56 ± 2.12	72.72 ± 1.29	26.87 ± 0.32
	Moco-v2	80.25 ± 0.93	71.85 ± 1.25	27.20 ± 0.66
	Barlow-Twins	80.95 ± 2.47	70.90 ± 1.89	26.83 ± 0.13
	Deep-Cluster-v2	81.03 ± 1.17	74.51 ± 1.92	28.03 ± 0.41
	SwAV	82.15 ± 1.19	74.50 ± 1.20	27.70 ± 0.22
2D & 3D Self-supervised	**Our (Deep-Cluster-v2)**	83.58 ± 1.54	78.14 ± 1.32	28.47 ± 0.40
	Our (SwAV)	84.89 ± 0.68	78.73 ± 1.21	27.47 ± 0.18

Table 5: Performance comparison on the 2D JSRT segmentation tasks using different SSL approaches trained on the *Unified setting*. Two top results are illustrated in red and blue respectively.

Pre-training	Methods	Backbone	JSRT (X-ray, seg.) 20%	40%	100%
N/A	Scratch CNN	ResNet-50	84.05	87.63	90.96
	Scratch PTU	Transformer	85.55	88.83	91.22
2D Supervised	Pre-trained ImNet	ResNet-50	87.90	90.01	91.73
2D Self-Supervised	Moco-v2	ResNet-50	88.65	91.03	92.32
	PGL	ResNet-50	89.01	91.39	92.76
	PCRL	ResNet-50	89.55	91.53	93.07
	Moco-v3	Transformer	90.07	91.75	92.68
	DINO	Transformer	90.40	92.16	93.03
2D & 3D Self-supervised	USST	Transformer	91.88	93.15	94.08
	Our (DeepCluster-V2)	ResNet-50	90.60	92.87	94.31
	Our (SwAV)	ResNet-50	89.98	93.03	94.45

We use ResNet-50 as the backbone for 2D feature extractor f_{intra}. The features for each image are concatenated from five blocks of ResNet-50. The architecture of f_{inter} has four pyramid structure blocks composed from deformable attention (Eq. (8)). Details for these configurations can be found in Appendix. In the *Universal* or *Unified* setting, we utilize all 3D data as benchmarks and further extract 2D slices from them to train f_{intra} in Stage 1. All experiments are conducted on a A100-GPU system with 4 GPUs, 40GB of memory each with Pytorch. It takes in average 30 hours to finish the pre-training step.

Downstream Task we use the SGD with a learning rate selected in a set $\{0.1, 0.01\}$ and select a specific number of epoch depended on downstream task properties (Appendix). The results are reported by running training-testing five times and computing the average values (except LUNA2016-v2 dataset). For the 2D/3D segmentation task, we use the pre-trained 2D-CNN feature extractor in each 2D baseline (f_{intra} in our method) as the network backbone of a 2D U-net (Ronneberger, Fischer, and Brox 2015). This network is trained with cross-entropy and dice loss. We predict segmentation at each 2D slice and merge results for 3D volumes. The 3D classification is solved by building on top of the deformable transformer two fully-connected layers and

fine-tuning for both f_{inter} and f_{intra} with the cross-entropy loss. For the 2D object detection task (VinDr-CXR), we use the 2D-CNN feature extractor (f_{intra}) as the backbone of Faster R-CNN model (Ren et al. 2015).

Performance Evaluation

Dimension-specific vs. Cross-dimension Pre-training Tables 3 and 4 indicate that 2D CNN based-models cannot transfer to the 3D lung node classification task in LUNA2016-v2 (denoted N/A) given input 3D volumes. Likewise, due to data compatibility issues, 3D CNN-based methods cannot apply for abnormal chest detection in X-rays. In contrast, our models pre-trained on several medical datasets can be transferred successfully in both cases due to the hybrid CNN-Transformer architecture. We argue that such property is one of the most valuable points of this study.

As compared with plain 2D-SwAV, Deepcluster-V2, and their extended versions with 3D CNN, namely 3D-SwAV and 3D-Deepcluster, we show a significant improvement in several settings, especially for segmentation tasks (Tables 3,4). For instance, a gain performance of 2-3% on average on BraTS, MMWHS-CT/MRI datasets. Furthermore, we also achieve better accuracy on 3D classification and 2D object detection, although with smaller margins. In conclusion, this analysis shows that exploiting deformable self-attention in conjunction with 2D CNN to model 3D volume features in our framework is a promising approach.

Comparison to SOTA Methods and Visualizations In the Universal setting, except the LUNA2016-v2 case where we are third, our methods based on Deepcluster-V2 or SwAV hold the best records on BraTS, MMWHS-CT/MRI segmentation tasks compared with remaining baselines, especially with cutting edges 3D-SSL methods as Universal Model, TransVW or SwinViT3D (using Swin Transformer). With the VinDr-CXR detection task, we continue to reach the best rank, followed by the plane 2D Deepcluster-v2 though with smaller margins. In the Unified setting, we also surpass competitors (100% data), especially with USST, a method using Pyramid Vision Transformer trained on mixed 2D and 3D data. However, USST works better than us when decreasing training data to 40% and 20%. We consider this as a potential limitation that needs to improve. Though it's worth noting that we could not access all data as USST in the pre-training step, as shown in Table 1.

For visualization results, we provide a typical example of multi-modal heart segmentation for MMWHS-CT in Figure 4 and abnormal Chest X-ray detection in Figure 5. More examples can be found in the Appendix.

Computational Complexity and Ablation Study We compare the total parameters with top baselines and methods using Transformer in Table 7. In short, our total parameter is half of the SwinViT3D but we attain better performance in overall. The contributions of proposed SSL tasks and multi-level features are presented in Table 6, where all components contribute to overall accurate growth.
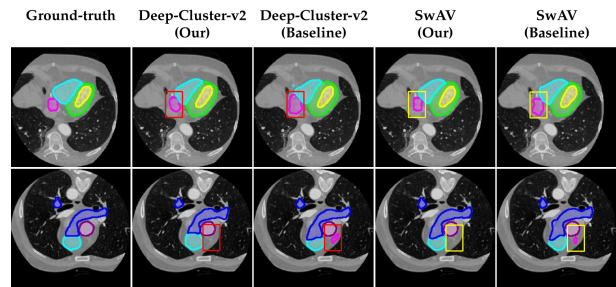

Figure 4: Heart structure segmentation on MMWHS-CT. The figures show that baselines tend to over-segment in the first row while generating noise regions in the second row. On the contrary, our methods produce more precise results.
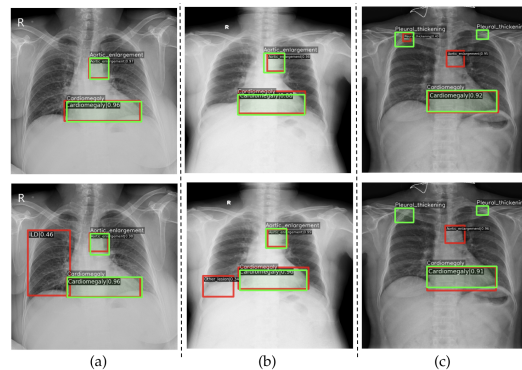

Figure 5: Visualization of our results based on DeepCluster-V2 (top row) and pre-trained ImageNet (second row) in abnormal Chest X-ray. Green and red indicates for ground-truths and predictions.

Table 6: Ablation studies for the SwAV on heart segmentation.

Setting	MMWHS - CT	MMWHS - MRI
W/o mask prediction	82.53	77.35
W/o 3D clustering	81.97	76.18
Full model	84.89	78.73
Full model w/o multi-feature	83.56	78.12

Table 7: Computational complexity of top baselines and transformer-based methods. For USST, we follow general descriptions in paper to re-configure architecture

.	SwinViT3D	TransVW	Universal Model	USST	Our
#Param	62.19 M	19.7 M	19.7 M	47.8 M	31.16 M

Conclusion

We contribute to the self-supervised learning medical imaging literature a new approach that is efficient in using numerous unlabeled data types and be flexible with data dimension barriers in downstream tasks. To that end, we developed a deformable self-attention mechanism on top of a 2D CNN architecture, which leads to both intra- and inter-correlations formed in our framework. Furthermore, our two novel SSL tasks including 3D agreement clustering and masked embedding predictions impose a tighter constraint in learning feature space, advancing pre-trained network performance in a variety of medical tasks. In the future, we will investigate this method for various SSL approaches, aiming to validate its universality and robustness in real-life medical usage.

Acknowledgement

Ethics Statement

Because our system does not require labeled data and benefits from as much data as possible, the temptation to exploit patient data without their consents arises. Ethical issues may also occur as a result of data collection bias which leads to unfavorable results for ethnically and economically disadvantaged subpopulations, thereby exacerbating the gaps currently existing in healthcare systems.

References

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. 2011. 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.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv:1811.02629*.

Benčević, M.; Habijan, M.; Galić, I.; and Pizurica, A. 2022. Self-Supervised Learning as a Means To Reduce the Need for Labeled Data in Medical Image Analysis. *arXiv preprint arXiv:2206.00344*.

Bilic, P.; Christ, P. F.; Vorontsov, E.; Chlebus, G.; Chen, H.; Dou, Q.; Fu, C.-W.; Han, X.; Heng, P.-A.; Hesser, J.; et al. 2019. The liver tumor segmentation benchmark (LiTS). *arXiv:1901.04056*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Neural Information Processing Systems (NeurIPS)*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33: 12546–12558.

Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; and Rueckert, D. 2019. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58: 101539.

Chen, S.; Ma, K.; and Zheng, Y. 2019. Med3d: Transfer learning for 3d medical image analysis. *arXiv:1904.00625*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Cheplygina, V.; de Bruijne, M.; and Pluim, J. P. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54: 280–296.

Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6): 1045–1057.

Cole, E.; Yang, X.; Wilber, K.; Mac Aodha, O.; and Belongie, S. 2022. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14755–14764.

de Leeuw den Bouter, M.; Ippolito, G.; O'Reilly, T.; Remis, R.; van Gijzen, M.; and Webb, A. 2022. Deep learning-based single image super-resolution for low-field MR brain images. *Scientific Reports*, 12(1): 1–10.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Do, K.; Tran, T.; and Venkatesh, S. 2021. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9928–9938.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ericsson, L.; Gouk, H.; and Hospedales, T. M. 2021. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5414–5423.

Gibson, E.; Li, W.; Sudre, C.; Fidon, L.; Shakir, D. I.; Wang, G.; Eaton-Rosen, Z.; Gray, R.; Doel, T.; Hu, Y.; et al. 2018. NiftyNet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158: 113–122.

Haghighi, F.; Taher, M. R. H.; Zhou, Z.; Gotway, M. B.; and Liang, J. 2021. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10): 2857–2868.

Harmon, S. A.; Sanford, T. H.; Xu, S.; Turkbey, E. B.; Roth, H.; Xu, Z.; Yang, D.; Myronenko, A.; Anderson, V.; Amalou, A.; et al. 2020. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature communications*, 11(1): 1–7.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, 574–584.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jin, L.; Yang, J.; Kuang, K.; Ni, B.; Gao, Y.; Sun, Y.; Gao, P.; Ma, W.; Tan, M.; Kang, H.; et al. 2020. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine*, 62: 103106.

Jun, E.; Jeong, S.; Heo, D.-W.; and Suk, H.-I. 2021. Medical transformer: Universal brain encoder for 3D MRI analysis. *arXiv preprint arXiv:2104.13633*.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2021. Prototypical contrastive learning of unsupervised representations. *International Conference on Learning Representations (ICLR)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Nguyen, D.; Nguyen, D. M.; Vu, H.; Nguyen, B. T.; Nunnari, F.; and Sonntag, D. 2021a. An attention mechanism using multiple knowledge sources for COVID-19 detection from CT images. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-2021), Workshop: Trustworthy AI for Healthcare*, volume 360.

Nguyen, D. H.; Nguyen, D. M.; Mai, T. T.; Nguyen, T.; Tran, K. T.; Nguyen, A. T.; Pham, B. T.; and Nguyen, B. T. 2022a. ASMCNN: An efficient brain extraction using active shape model and convolutional neural networks. *Information Sciences*, 591: 25–48.

Nguyen, D. M.; Mai, T. T.; Than, N. T.; Prange, A.; and Sonntag, D. 2021b. Self-supervised Domain Adaptation for Diabetic Retinopathy Grading Using Vessel Image Reconstruction. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, 349–361. Springer.

Nguyen, D. M.; Nguyen, T. T.; Vu, H.; Pham, Q.; Nguyen, M.-D.; Nguyen, B. T.; and Sonntag, D. 2022b. TATL: task agnostic transfer learning for skin attributes detection. *Medical Image Analysis*, 102359.

Nguyen, D. M.; Vu, H. T.; Ung, H. Q.; and Nguyen, B. T. 2017. 3D-brain segmentation using deep neural network and Gaussian mixture model. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 815–824. IEEE.

Nguyen, D. M. H.; Ezema, A.; Nunnari, F.; and Sonntag, D. 2020. A visually explainable learning system for skin lesion detection using multiscale input with attention U-Net. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, 313–319. Springer.

Nguyen, H. Q.; Lam, K.; Le, L. T.; Pham, H. H.; Tran, D. Q.; Nguyen, D. B.; Le, D. D.; Pham, C. M.; Tong, H. T.; Dinh, D. H.; et al. 2022c. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1): 1–7.

Pan, E.; and Kang, Z. 2021. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34: 2148–2159.

Raghu, M.; Zhang, C.; Kleinberg, J.; and Bengio, S. 2019. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Setio, A. A.; Jacobs, C.; Gelderblom, J.; and van Ginneken, B. 2015. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical physics*, 42(10): 5642–5653.

Shiraishi, J.; Katsuragawa, S.; Ikezoe, J.; Matsumoto, T.; Kobayashi, T.; Komatsu, K.-i.; Matsui, M.; Fujita, H.; Kodera, Y.; and Doi, K. 2000. 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. *American Journal of Roentgenology*, 174(1): 71–74.

Simpson, A. L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; Van Ginneken, B.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:1902.09063*.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Taleb, A.; Loetzsch, W.; Danz, N.; Severin, J.; Gaertner, T.; Bergner, B.; and Lippert, C. 2020. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33: 18158–18172.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740.

Van Ginneken, B.; Stegmann, M. B.; and Loog, M. 2006. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1): 19–40.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, D.; Zhang, Y.; Zhang, K.; and Wang, L. 2020. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3951–3960.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794–4803.

Xie, Y.; Zhang, J.; Liao, Z.; Xia, Y.; and Shen, C. 2020. PGL: prior-guided local self-supervised learning for 3D medical image segmentation. *arXiv preprint arXiv:2011.12640*.

Xie, Y.; Zhang, J.; Xia, Y.; and Wu, Q. 2021. Unified 2D and 3D Pre-training for Medical Image classification and Segmentation. *arXiv preprint arXiv:2112.09356*.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230*.

Zhang, J.; Xie, Y.; Xia, Y.; and Shen, C. 2021. DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1195–1204.

Zhang, X.; Zhang, Y.; Zhang, X.; and Wang, Y. 2020. Universal Model for 3D Medical Image Analysis. *CoRR*, abs/2010.06107.

Zhou, H.-Y.; Lu, C.; Yang, S.; Han, X.; and Yu, Y. 2021a. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3499–3509.

Zhou, Z.; Sodha, V.; Pang, J.; Gotway, M. B.; and Liang, J. 2021b. Models genesis. *Medical image analysis*, 67: 101840.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhuang, X.; Li, Y.; Hu, Y.; Ma, K.; Yang, Y.; and Zheng, Y. 2019. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 420–428. Springer.

Zhuang, X.; and Shen, J. 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31: 77–87.

Appendix – Joint Self-Supervised Image-Volume Representation Learning with Intra-Inter Contrastive Clustering

In this supplement material, we present more information on our deformable self-attention architecture, pre-training, downstream setup, an ablation study when changing the rate of training samples in fine-tuning phases, additional visualization illustrations, and finally, discuss current limitations and future directions for further improvements.

Deformable Self-attention Architecture

To construct a hierarchical feature representation, we build the deformable self-attention with four consecutive blocks as Figure 6. Each block involves 6 number of heads and 6 offset groups (Zhu et al. 2020; Xia et al. 2022). Intuitively, this design permits employing a multi-scale information of consecutive 2D slices in 3D volumes, thereby enhancing representation capability while saving computational costs, especially for 3D classification task when attention is estimated for selected local regions.

Pre-Training Setting

2D Self-Supervised Baselines: We employ VISSL[1] to perform pre-training for all 2D self-supervised methods based on ResNet-50 backbone, including SimCLR, Moco-v2, Barlow-Twins, Deepcluster-v2, and SwAV. All of them are trained with 100 epochs, batch size of 1024 images distributed on 4 GPUs of the A100 system. Other parameters are set as default by VISSL. Training data for these 2D-SSL approaches is taken by extracting 2D slices from all 3D volumes in distinct datasets in the *Universal* setting. The optimizes and learning rates are followed as standard 2D Deepcluster-v2 and SwAV.

3D Self-Supervised Baselines: For two 3D SSL methods, 3D-Deepcluster and 3D-SwAV, we implemented the 3D variant of Deepcluster-v2 and SwAV by using 3D-CNN, then we trained them with all 3D data in the *Universal* setting using a batch size of 12 volumes distributed on 4 GPUs of A100 system with 100 epochs. The optimizer, learning rate, and other configurations are the same as in the 2D case.

Our Method: We optimize the proposed method in three stages. Stage 1 learns f_{intra} with ResNet-50 using Eq. (5) with 100 epochs, i.e., training default Deepcluster-v2 or SwAV. Stage 2 learns f_{inter} and f_{decode} using Eq.(15) with 100 epochs given a batch size of 12 volumes. Finally, we jointly optimize f_{intra} and f_{inter} using the novel 3D agreement clustering problem defined in Eq. (12) with $100/200$ epochs in Universal/Unified settings. Due to the expensive computational costs inside the Deepcluster-v2 compared to SwAV, we specified a batch size of 8 for this method to avoid out-of-memory issues. With SwAV, we used a batch size of 24 and selected the number of hidden clusters $H = 3000$.

---

[1] https://github.com/facebookresearch/vissl

We leverage both 2D images and 3D data. In the *Universal* setting, since the data is only available in 3D formats, we get all 2D slices from them to train f_{intra} in Stage 1 as other 2D SSL baselines. Afterward, Stages 2 and 3 use 3D data as usual. Similarly, in the *Unified* setting, we also get 2D images from 3D CT and merge them with 2D X-rays in the NIH ChestX-ray8 dataset to train Stage 1. Similarly, stages 2 and 3 only use all available 3D CT as 3D SSL methods.

Dataset Description and Other Settings in Downstream Tasks

Downstream Dataset

We briefly provide information on datasets used in downstream task in Table 8. This includes training, validation, testing sizes, and the employed loss functions in corresponding tasks. Below we describe more details each dataset's properties:

- *BraTS2018* (Bakas et al. 2018): This dataset comprises magnetic resonance imaging (MRI) volumes of 285 brain tumor patients. Each participant was scanned using four distinct modalities: T1-weighted, T1-weighted with contrast enhancement, T2-weighted, and T2 fluid-attenuated inversion recovery (FLAIR). The voxel-level labels of the "whole tumor", "tumor core", and "enhancing tumor" are annotated for each patient. Following settings of baselines in (Zhang et al. 2020), we choose FLAIR images and build model for the "whole tumor". The training and testing rates are indicated in Table 8.

- *LUNA2016* (Setio et al. 2015): This dataset consists of 888 computed tomography (CT) scans, aiming for nodule detection and false positive reduction task. We employed the extended set including 754975 candidates extracted from 888 CT scans, which is 203910 candidates more than the original set (denoted as LUNA2016-v2). The training set has 817 positive lung nodules out of 377138 detection candidates. There are 225475 candidates in the test set, including 459 positive lung nodules. The average shape for each candidate is $64 \times 64 \times 32$.

- *MMWHS-CT/MRI* (Zhuang and Shen 2016): This dataset is made up of unpaired 20 MRI and 20 CT volumes that span the whole heart substructures and includes seven labeled structures. Following baseline settings (Zhang et al. 2020), we segment the left atrial blood cavity regions on the CT and MR formats.

- *VinDr-CXR* (Nguyen et al. 2022c): This dataset aims to localize organs and nodules from 2D X-ray lung images. The classes include aortic enlargement, atelectasis, calcification, cardiomegaly, consolidation, ILD, infiltration, lung opacity, nodule/mass, pleural effusion, pleural thickening, pneumothorax, pulmonary fibrosis, and other lesions. The total of X-ray images has abnormal tissues is 4394 images. For images where labels of the same class
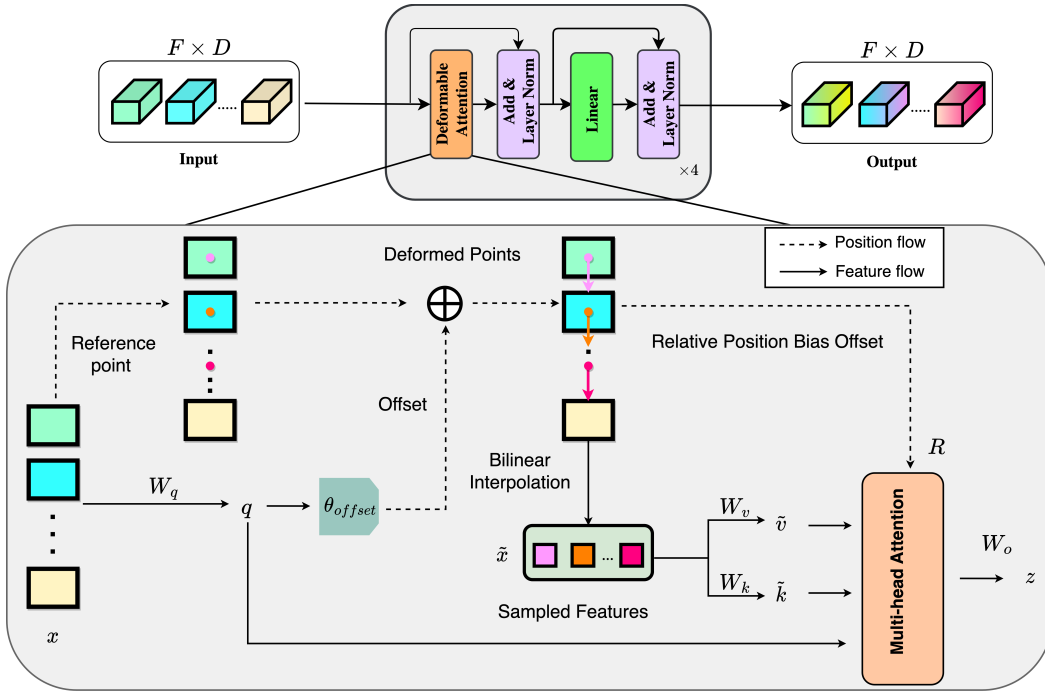
Figure 6: The illustration of the Deformable Attention Transformer. F refers to the number of frames of our 3D input and D represents the embedding dimension for each frame. In our setting, we choose $F = 64$, $D = 640$.

obtained from different experts are overlapped, we pre-process by averaging overlapping bounding boxes with an intersection-over-union of 20%. The training, validation, and testing rate are presented in Table 8. In this task, we build a model to detect all nodules available in testing images.

- *JSRT* (Shiraishi et al. 2000; Van Ginneken, Stegmann, and Loog 2006): This dataset includes 2D X-ray images taken from lung organs. The annotations consist of the heart, left clavicle, right clavicle, left lung, and right lung. We construct models to segment all organs in testing images followed by baselines in (Xie et al. 2021).

Other Settings

2D & 3D Segmentation Tasks: We formulate 3D segmentation tasks in BraTS2018, MMWHS-CT/MRI as the 2D segmentation problem (JSRT data). To this end, we create two subsets to avoid imbalance issues. The first one includes 2D images whose labels contain a target object required to segment, and the other comprises background and remaining structures. We then sample data from these two sets and train with a U-Net model. The backbone for this U-Net is the pre-trained ResNet-50 (f_{intra} in our model). We employ the SGD as the optimizer and learning rate 0.1 for segmentation-related tasks. All 2D SSL baselines are trained with 50 epochs. Our method usually converges faster so we picked $15, 25, 50, 50$ epochs for MMWHS-MRI, MMWHS-CT, BraTS2018, and JSST (100% case) respectively. The results are reported in average performance over five trial times using 3D Intersection over Union (IoU) for 3D settings and 2D Dice for 2D cases.

3D Object classification (LUNA2016-v2) Given the output feature maps of deformable attention, we build on top of the f_{intra} and f_{inter} two fully connected layers in the size of 640 and 252. The last layer returns two probability outputs for the binary classification problem. We use the binary cross entropy as the loss function and train with SGD using learning rate of 0.1 for 100 epochs. The Area Under the Curve (AUC) metric is used to evaluate performance. We report average performance for different batch sizes of $8, 16, 32$ each case with two trials.

2D Object Detection (VinDr-CXR) We choose the Faster R-CNN from MMDetection[2] framework as a base model for the 2D object detection. This model is loaded pre-trained ResNet-50 derived from different 2D SSL methods. With 2D SSL baselines, the detector is trained with SGD as the optimizer with a learning rate $lr = 0.001$ and converges after 25 epochs. However, our pre-trained weights could not train with this learning rate (NAN loss), we thus selected $lr = 0.1$ for the first 30 epochs and rescheduled $lr = 0.01$ for the last 10 epochs to get stable checkpoints (normally converge after 35 epochs). Other parameters are set as default by MMDetection. We use the mean average precision (mAP) of all classes with an IoU threshold fixed at 0.5 (mAP@0.5) for the evaluation as (Benčević et al. 2022). The results are computed in average of three trial times.

---

[2]https://github.com/open-mmlab/mmdetection

Table 8: Details of datasets for downstream tasks. For loss functions, *Dice* indicates dice loss, *CE* indicates cross-entropy loss, and *L1* indicates L1-loss.

Dataset	MMWHS-CT	MMWHS-MRI	BraTS2018	LUNA2016-v2	VinDr-CXR	JSRT
Task	3D Segmentation	3D Segmentation	3D Segmentation	3D Classification	2D object detection	2D Segmentation
Modality	3D CT	3D MRI	3D MRI	3D CT	2D X-ray	2D X-ray
Training samples	13	13	133	377138	3075	114
Validation samples	3	3	57	152362	440	10
Test samples	4	4	95	225475	879	123
Mean Data Size	$224 \times 224 \times 265$	$224 \times 224 \times 145$	$224 \times 224 \times 155$	$64 \times 64 \times 32$	512×512	224×224
Loss	Dice + CE	Dice + CE	Dice + CE	CE	L1 + CE	Dice + CE

Performance of Pre-trained Models when Varying Training Data

In this experiment, we investigate behaviors of our pre-trained model (extended from Deepcluster-v2) on downstream tasks when varying training data size. We conduct testings on the VinDr-CXR dataset by increasing training samples from 10% to 100% during the fine-tuning phase. Figure 7 indicates class-wise average precision and the overall performance across classes (dashed black curve). As can be seen, there exists a trend of improving performance when more data is available; however, at the rate of 80%, we already achieved a comparable accuracy of 90% or 100%. This evidence suggests that the proposed SSL method can use fewer labeled data but still can achieve similar performance. We argue that this property is valuable, especially in medical applications where obtaining labeled instances is extremely expensive.
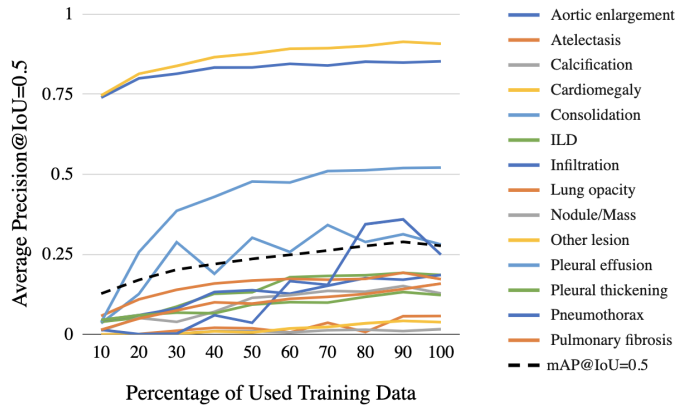
Figure 7: The mean average precision across different percentages of labeled data in abnormal Chest X-rays detection.

More Visualization Results

Figure 8 provides qualitative results on detecting abnormal nodules 2D X-rays in the VinDr-CXR. We compare our extended version using Deepcluster-v2 with the pre-trained Imagenet. Across different testing images, it seems that the proposed method might help reduce false positive predictions. For instance, the baseline incorrectly signifies the pleural thickening structure in the first and second columns. In the third case, though our method also wrongly predicts pleural thickening, the errors are less than the baseline when

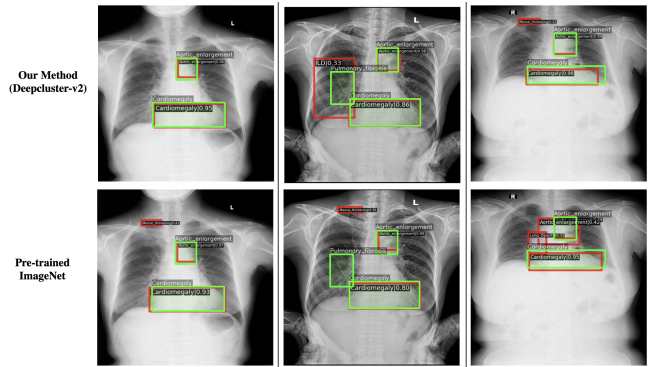there are wrong bounding boxes for either the lung opacity or aortic enlargement.

Figure 8: Qualitative results on detecting abnormal nodules in the VinDr-CXR dataset. Our method uses Deepcluster-v2, and the pre-trained Imagenet backbone is the first and second row. Green and red indicate ground truths and predictions, respectively.

Current Limitations and Future Directions

While the proposed method achieves state-of-the-art performance in several settings, the results on 2D object detection (VinDr-CXR) and 3D classification tasks (LUNA2016-v2) have not improved with large margins or still be smaller compared to other 3D SSL competitors. We argue that our learned features probably do not aid significantly these downstream tasks. We believe that further improvements might be gained by integrating downstream tasks' properties in the pre-training algorithms as current insights in (Ericsson, Gouk, and Hospedales 2021; Cole et al. 2022).

We also recommend conducting additional experiments on color images, such as skin attribute segmentation/classification (Nguyen et al. 2020; Sun et al. 2021), diabetic retinopathy grading (Nguyen et al. 2021b; Sun et al. 2021), or low-resolution images (Nguyen et al. 2017; de Leeuw den Bouter et al. 2022), to further validate the method's generalizations. Furthermore, it is critical to extending the framework with a similar mechanism for other SSL methods or learning under scenarios such as data imbalance or domain shift.