Cross-Dimensional Medical Self-Supervised Representation Learning Based on a Pseudo-3D Transformation

Fei Gao[1*], Siwen Wang[2*], Fandong Zhang[2], Hong-Yu Zhou[3], Yizhou Wang[4,5], Churan Wang[1†], Gang Yu[6†], and Yizhou Yu[7†]

[1] School of Computer Science, Peking University, Beijing, China
[2] Deepwise AI Lab, Beijing, China
[3] Department of Biomedical Informatics, Harvard Medical School, Boston, USA
[4] CFCS, School of Computer Science, Peking University, Beijing, China
[5] Institute for Artificial Intelligence, Peking University, Beijing, China
[6] Children's Hospital of Zhejiang University School of Medicine, Hangzhou, China
[7] Department of Computer Science, The University of Hong Kong, Hong Kong

Abstract. Medical image analysis suffers from a shortage of data, whether annotated or not. This becomes even more pronounced when it comes to 3D medical images. Self-Supervised Learning (SSL) can partially ease this situation by using unlabeled data. However, most existing SSL methods can only make use of data in a single dimensionality (e.g. 2D or 3D), and are incapable of enlarging the training dataset by using data with differing dimensionalities jointly. In this paper, we propose a new cross-dimensional SSL framework based on a pseudo-3D transformation (CDSSL-P3D), that can leverage both 2D and 3D data for joint pre-training. Specifically, we introduce an image transformation based on the im2col algorithm, which converts 2D images into a format consistent with 3D data. This transformation enables seamless integration of 2D and 3D data, and facilitates cross-dimensional self-supervised learning for 3D medical image analysis. We run extensive experiments on 13 downstream tasks, including 2D and 3D classification and segmentation. The results indicate that our CDSSL-P3D achieves superior performance, outperforming other advanced SSL methods.

Keywords: Self-supervised Learning · Pseudo-3D transformation · Medical image analysis.

1 Introduction

Medical image analysis often suffers from the lack of high-quality annotated data, which hinders the development of this field. This is primarily due to the labor-intensive and time-consuming nature of data annotation, especially for 3D data such as CT and MRI scans. Recently, self-supervised learning (SSL) has

*Equal contributions.
†Corresponding authors.

emerged as a promising approach to reduce the demand for annotated data by leveraging unlabeled data for representation learning [4, 7, 8, 18, 25, 26].

However, most existing self-supervised methods are typically confined to training on either 2D or 3D data exclusively, due to the dimensionality disparity. The integration of 2D and 3D data for joint self-supervised training could substantially increase the amount of pre-training data, potentially enhancing the quality of representation learning for 3D medical image analysis. There have been a few efforts to tackle this challenge. UniMiSS [23] proposes the adoption of a switchable patch embedding module to accommodate both 2D and 3D inputs. Nevertheless, this approach is only applicable to transformers and not compatible with CNN models. Note that CNN is also a powerful neural architecture, widely used in a variety of medical image analysis applications. Nguyen et al. [16] proposed to integrate 2D CNNs with deformable attention transformers, which can simultaneously extract 2D and 3D features. However, this approach results in a disjoint representation of 2D and 3D features, and the overall rigid architecture is not compatible with existing optimized neural architectures.

To tackle the aforementioned issues, we propose a **C**ross-**D**imensional **S**elf-**S**upervised **L**earning framework based on a **P**seudo-**3D** transformation, referred to as **CDSSL-P3D**. This framework overcomes the limitations imposed by dimensional disparity and is not confined to specific neural architectures, making it a genuinely cross-dimensional SSL strategy. Specifically, drawing inspiration from the im2col [5, 19] technique employed in convolution computations, we transform 2D images by sliding a window across them and unfolding the regions within each window into columns. This approach enables us to treat 2D images as 3D data. After this transformation, both 2D and 3D images can be concurrently fed into a neural network without necessitating any modifications to the architecture itself. Consequently, this seamless integration permits the direct application of existing SSL methods for the purpose of pre-training a 3D model. In our experiments, we adopt the pretext tasks proposed by [26], which preserves both pixelwise and semantic information in representation. We conduct model pre-training on a dataset comprising 6,453 3D volumes and 377,088 X-ray images. With the inclusion of the substantial collection of X-ray images into the training set, there is a noticeable improvement in performance for downstream tasks of 3D classification and segmentation. As an additional benefit, performance on 2D classification tasks can also be improved.

Overall, our contributions in this paper can be summarized as follows: (1) We propose a novel approach (CDSSL-P3D) based on the im2col transformation to tackle the challenge imposed by joint self-supervised pre-training using both 2D and 3D data, making SSL more flexible with respect to image dimensionality. (2) Our proposed method is compatible with the full spectrum of CNN and transformer architectures, not restricted solely to a specific neural architecture. (3) Our CDSSL-P3D method achieves significant performance improvements across 13 downstream tasks, including 3D medical image classification and segmentation as well as 2D medical image classification.

2 Method

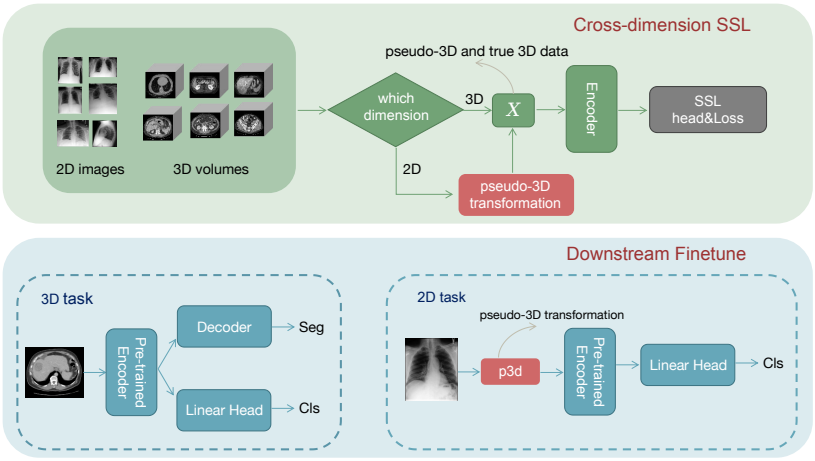

Fig. 1: The overall CDSSL-P3D framework. In the pre-training stage, 2D images are converted to pseudo-3D images. Then, SSL is performed on the joint pseudo-3D and true 3D data. During the fine-tuning stage, this pre-trained 3D model is primarily used for downstream 3D tasks. As an additional benefit, downstream 2D classification tasks can be supported, and images in such 2D tasks go through our pseudo-3D transformation before fed into the 3D model.

The overall process is shown in Fig. 1. During the pre-training phase, we initially employ a pseudo-3D transformation to convert all 2D images into a 3D format. Subsequently, we conduct self-supervised training of a 3D model using both genuine 3D images and the transformed pseudo-3D images. This approach enables cross-dimensional representation learning. This pre-trained 3D model can be applied to 3D downstream tasks as usual. As an additional benefit, 2D classification tasks can also be carried out by converting 2D images to 3D format first using our pseudo-3D transformation. Inspired by the im2col transformation used in convolution operators, we propose a pseudo-3D transformation in a similar manner to convert 2D images into a 3D format.

2.1 Preliminary: image-to-column transformation (im2col)

The application of the im2col [5,19] method has been thoroughly investigated for its effectiveness in transforming the Multiple Channel Multiple Kernel (MCMK) problem into the framework of General Matrix Multiplication (GEMM). It is widely used to accelerate convolution computation. As shown in Fig 2, assume an input $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$ and M kernels $\mathcal{K} \in \mathbb{R}^{M \times k \times k \times C}$. From the input $\mathcal{I}$ we could construct a new *input-patch-matrix* $\widehat{\mathcal{I}}$ by copying *patches* out of the input and unrolling them into columns of $\widehat{\mathcal{I}}$. These patches are formed in the shape of the kernel (i.e. $k \times k \times C$) at every location in the input where the kernel is to

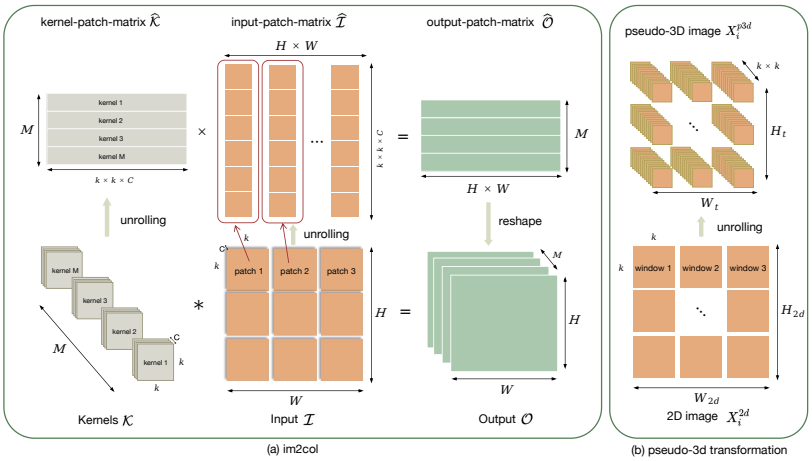

Fig. 2: The proposed pseudo-3D transformation inspired by im2col for MCMK problem. (a) Detailed depiction of im2col. Input image $\mathcal{I}$ and convolution kernel $\mathcal{K}$ are first unrolled into matrices $\widehat{\mathcal{I}}$ and $\widehat{\mathcal{K}}$, which are then multiplied to obtain the output. (b) Pseudo-3D transformation. Inspired by the transformation of $\widehat{\mathcal{I}}$, every instance of a sliding window over the entire 2D image X_i^{2d} is unrolled to obtain the pseudo-3D image X_i^{p3d}.

be applied. Afterwards, the dimension of the transformed input-patch-matrix $\widehat{\mathcal{I}}$ will be $H_p \times W_p$ as:

$$H_p = k \times k \times C \tag{1}$$

$$W_p = (\frac{H + 2P - k}{s} + 1) \cdot (\frac{W + 2P - k}{s} + 1) \tag{2}$$

where P and s are padding and stride in convolution.

Once the input-patch-matrix $\widehat{\mathcal{I}}$ is formed, we construct the kernel-patch-matrix $\widehat{\mathcal{K}}$ by unrolling each of the M kernels of shape $k \times k \times C$ into one row of $\widehat{\mathcal{K}}$. The shape of the resulting matrix $\widehat{\mathcal{K}}$ is $M \times (k \times k \times C)$. Then we simply perform a GEMM between $\widehat{\mathcal{K}}$ and $\widehat{\mathcal{I}}$ to obtain the output $\widehat{\mathcal{O}} \in \mathbb{R}^{H \times W \times M}$.

2.2 Pseudo-3D transformation based on im2col

Notation: We assume to be given 2D and 3D datasets $\{\mathcal{D}^{2d}, \mathcal{D}^{3d}\}$ where $\mathcal{D}^{2d} = \{X_i^{2d}\}, i \in [1, N^{2d}]$ and $\mathcal{D}^{3d} = \{X_i^{3d}\}, i \in [1, N^{3d}]$. Denote 2D image $X_i^{2d} \in \mathbb{R}^{H_{2d} \times W_{2d}}$ and 3D volume $X_i^{3d} \in \mathbb{R}^{H_{3d} \times W_{3d} \times D_{3d}}$.

Intuitively, transforming input image $\mathcal{I}$ into input-patch-matrix $\widehat{\mathcal{I}}$ via im2col motivates us to conceive a pseudo-3D transformation on 2D images. Specifically, given a window size $k \times k$ and stride s, similar to a convolution kernel in im2col, a 2D image X_i^{2d} can be transformed to $X_i^{p3d} \in \mathbb{R}^{H_t \times W_t \times D_t}$, where

$$H_t = \frac{H_{2d} - k}{s} + 1, W_t = \frac{W_{2d} - k}{s} + 1, D_t = k \times k. \tag{3}$$

This transformation is slightly different from construction of input-patch-matrix $\widehat{\mathcal{I}}$ in im2col. In our strategy, the windows across rows and columns of X_i^{2d} are maintained in two dimension of H_t and W_t, instead of one dimension as W_p (Eq. 2). Through such pseudo-3D transformation, the original information in 2D space is converted to 3D space. Such transformed data can be potentially beneficial to capture complex 3D structure and texture representations in 3D medical images for any 3D model. Using the proposed pseudo-3D transformation, we are capable of generating large-scale 3D datasets that significantly surpass the scale of existing publicly available 3D medical data. And, any 3D network can be trained concurrently on both pseudo-3D data and true 3D volume data, ensuring a seamless integration and finally a cross-dimensional SSL framework.

2.3 Learning Objective

The essence of this paper lies in cross-dimensional self-supervised learning, thus it is not confined to any specific self-supervised training strategy; in principle, all existing strategies are viable. We have opted for the relatively recent and powerful approach PCRLv2 [26] as an example in this study. PCRLv2 addresses information preservation in self-supervised visual representations from three aspects: pixels, semantics, and scales. First, a pixel-level objective of reconstructing the precise pixel-level details from corrupted inputs could force the model to capture pixel information in feature representations. Second, high-level siamese feature comparison is adopted to preserve semantic information in latent representations. In addition, multi-scale reconstruction and feature comparison are conducted to learn multi-scale representations.

2.4 Network

Our approach is not constrained to any specific neural architecture and is compatible with both prevailing architectures, CNNs and Transformers. For CNN, we use a 3D version of ResNet-18 [9] as the encoder, a commonly used and efficient network. In the pre-training stage, a U-like architecture with the encoder stacked with a CNN-based decoder is adopted. Regarding transformer, we adopt the pyramid vision transformer (PVT) designed for large-scale vision tasks [20], which is also adopted in [23] for cross-dimensional SSL. Unlike that in [23], our patch embedding strategy does not necessitate switching based on data dimensions, resulting in a unified and simplified model structure. Specifically, we conduct experiments with PVT-small in this study.

3 Experiments

3.1 Datasets

Pre-training datasets. We collect 6,453 3D CT and MRI volumes from eight public datasets (LUNA16 [17], RibFrac [12], TCIA Covid19 [1], AMOS22 [11], ISLES2022 [10], AbdomenCT-1K [15], Totalsegmentor [22], Verse 2020 [14]) and 377,088 2D X-ray images from the MIMIC-CXR dataset [13] for cross-dimensional self-supervised learning.

Downstream datasets. To thoroughly evaluate the effectiveness of the pre-training, we conducted comparative experiments across 13 downstream tasks. These tasks can be categorized into the following groups: (1) 3D classification (MedMNIST v2 [24], including six individual 3D tasks of various medical images), (2) 3D segmentation (comprising six datasets of the Liver, Hepatic Vessel (HepaV), Pancreas, Colon, Lung, and Spleen dataset from Medical Segmentation Decathlon (MSD) [2]), (3) 2D classification (NIH ChestX-ray) [21].

Table 1: Classification results of AUC on the test sets of the six 3D image datasets from MedMNIST v2 [24]. The results of ResNet-18+3D is taken from the original paper [24]. The best results of each backbone are bolded and the second-best are underlined.

Method	Backbone	organ	nodule	fracture	adrenal	vessel	synapse	average
ResNet-18+3D [24]	ResNet-18	0.996	0.863	0.712	0.827	0.874	0.820	0.848
DINO [3]		0.995	0.890	0.707	0.847	0.918	0.810	0.861
SimSiam [6]		0.995	0.874	0.738	0.837	0.876	0.765	0.848
TransVW [8]		**0.998**	0.898	0.731	0.835	0.905	0.811	0.863
PCRLv2 [26]		0.995	0.894	0.740	0.853	0.930	0.798	0.868
CDSSL-P3D		**0.998**	**0.908**	**0.754**	**0.880**	**0.947**	**0.835**	**0.888**
Rand. init.	PVT-small	0.980	0.876	0.651	0.824	0.907	0.770	0.835
UniMiSS [23]		**0.996**	0.894	0.724	0.853	0.927	0.847	0.874
CDSSL-P3D		0.993	**0.930**	**0.761**	**0.857**	**0.960**	**0.912**	**0.902**

3.2 Experimental Details

Pre-training setup. For the 2D MIMIC-CXR dataset, each image is resized to 224×224 after random crop and then transformed as a pseudo-3D patch. For 3D datasets, we randomly crop a patch from the whole CT volume with size from $\{64 \times 64 \times 32, 96 \times 96 \times 48, 112 \times 112 \times 56, 128 \times 128 \times 64\}$. The cropped patches are then resized to $64 \times 64 \times 32$. The input patch size is determined to strike a balance between preserving sufficient information for SSL and lower computational complexity to a manageable level. As in [26], for a given input image, a two-stage augmentation strategy is performed to corrupted it in global and local aspects. Global augmentation includes random flip and random affine. Local augmentation includes random noise, Gaussian blur, random swap, and random gamma. We employ Adam as the default optimizer and a learning rate with cosine decaying initial from 1e-3. The epochs of training is empirically set to 200, with a batch size of 96.

Downstream training setup. For all the downstream tasks, only encoder is initialized from the pre-trained models. For 3D classification tasks within MedMNIST v2, the official test set is adopted to evaluate the models performance, and the performance is measured by area under the receiver operator curve (AUC). Regarding 3D segmentation tasks, we randomly split the data of each task into training, validation and test at a ratio of 7:1:2. The Dice score is employed as

evaluation metric. For 2D classification of NIH ChestX-ray, the training, validation, test sets are also randomly divided as 7:1:2, and similarly, AUC is used as the performance metric.

Table 2: Quantitative results on six 3D segmentation datasets. We compare the Dice (%) on each dataset and average Dice (%) of all datasets. The best results of each backbone are highlighted in bold and the second-best are underlined.

Method	Backbone	Liver	HepaV	Pancreas	Colon	Lung	Spleen	average
Rand. init.	ResNet-18	75.2	60.3	60.1	30.6	42.2	92.1	60.1
DINO [3]		76.0	60.8	61.3	37.5	45.6	92.0	62.2
SimSiam [6]		76.6	62.3	61.2	32.5	46.2	92.0	61.8
TransVW [8]		76.9	61.1	61.9	32.7	46.5	93.4	62.1
DeSD [25]		76.8	62.2	61.8	40.2	52.5	93.9	64.6
PCRLv2 [26]		79.3	62.3	61.5	43.2	54.2	95.6	66.0
vox2vec [7]		78.5	63.8	61.8	32.6	47.2	96.1	63.3
CDSSL-P3D		**81.2**	**65.0**	**63.0**	**46.2**	**57.1**	**96.2**	**68.1**
Rand. init.	PVT-small	77.9	63.6	63.5	39.3	49.3	93.5	64.5
UniMiSS [23]		81.1	64.3	64.1	44.6	56.7	95.4	67.7
CDSSL-P3D		**82.5**	**67.8**	**65.5**	**50.4**	**60.5**	**96.2**	**70.5**

3.3 Results

Comparing to other SSL Methods. The proposed CDSSL-P3D is compared with random initialization, and seven advanced SSL methods including DINO [3], SimSiam [6], TransVW [8], DeSD [25], PCRLv2 [26], vox2vec [7] and UniMiSS [23]. Note that the first six methods use CNNs as their encoder backbone and UniMiSS [23] adopts a transformer as its backbone. In addition, only UniMiSS exploits cross-dimensional data as ours while the first six methods use data with a single dimensionality only. Thus, in the comparison experiments, the first six methods are pre-trained using data with the same dimensionality as the downstream tasks. Our CDSSL-P3D and UniMiSS are pre-trained on all 2D and 3D data collected for pre-training. As detailed in Tables 1, 2, the proposed CDSSL-P3D is compared with the competitors primarily on 3D medical tasks including six 3D classification tasks (Table 1) and six 3D segmentation tasks (Table 2). In addition, one 2D classification task (Table 3) is also conducted because downstream 2D classification tasks are supported by our pre-trained 3D model. The following conclusions can be drawn from the tables: 1) SSL significantly enhances model performance compared with random initialization. 2) Transformer-based models generally outperform CNN-based methods. 3) Our CDSSL-P3D framework demonstrates notable performance improvements for both CNNs and Transformers, confirming the effectiveness of our cross-dimensional strategy on the two predominant neural architectures. 4) CDSSL-P3D achieves the highest performance across all tasks, surpassing the second-best methods, PCRLv2 [26] and UniMiSS [23], by 2.0%, 2.8% (3D classification), 2.1%, 2.7% (3D segmentation) and 1.0%, 1.6% (2D classification), respectively. Note that the performance does

not deteriorate with the 3D model compared with the 2D model on the NIH ChestX-ray dataset, indicating that employing pseudo-3D transformations for 2D downstream tasks does not incur a loss in performance (first and sixth rows in Table 3). Furthermore, given the substantial size of the NIH ChestX-ray dataset (over 100,000 images), we conduct additional experimental comparisons at varying training data ratios (Table 3). The results indicate that the CDSSL-P3D framework consistently provides the most significant improvement across different ratios, with particularly notable improvements at smaller training set sizes.

Table 3: Quantitative results of different SSL strategies on NIH ChestX-ray dataset for 2D Classification, measured by AUC under different ratios of training data. The best results are bolded and the second-best are underlined.

Method	Backbone	10%	30%	50%	100%
Rand. init.	ResNet-18 2D	0.695	0.735	0.774	0.808
DINO [3]		0.723	0.790	0.787	0.818
SimSiam [6]		0.728	0.785	0.789	0.819
TransVW [8]		0.715	0.753	0.788	0.816
PCRLv2 [26]		0.770	0.809	0.819	0.828
Rand. init.	ResNet-18 3D	0.703	0.745	0.772	0.805
CDSSL-P3D		**0.778**	**0.818**	**0.827**	**0.838**
Rand. init.	PVT-small	0.712	0.764	0.782	0.816
UniMiSS [23]		0.771	0.809	0.820	0.840
CDSSL-P3D		**0.789**	**0.823**	**0.840**	**0.856**

Ablation of pre-training with different dimension. A key contribution of this paper is the joint pre-training of 2D and 3D data, which offers distinct advantages over pre-training with data from a single dimension alone. To substantiate the efficacy of joint pre-training, we have compared the models performance pre-training with different data dimensions (2D, 3D, 2D+3D) on downstream 2D (NIH ChestX-ray) and 3D (MedMNIST v2) tasks (shorten as NIH and MedM in Table 4, 5, 6). This comparison is conducted on ResNet-18 as a example, which is also adopted in the following ablation studies. As depicted in Table 4, joint pre-training exhibits a significant enhancement compared to using either 2D or 3D data exclusively.

Ablation on window size. Table 5 presents the results under various window sizes (3×3, 5×5, 7×7). Overall, larger windows tend to yield superior performance (5×5 and 7×7 outperform 3×3). Nevertheless, it is not the case that larger windows always lead to better results. The optimal performance is achieved with a 5x5 window size. We speculate that the advantage of larger window sizes may be attributed to the compatibility with the dimensionality of 3D data, leading to better joint pre-training integration.

Table 4: Ablation of SSL dimension.

dim	MedM	NIH
2D	0.863	0.828
3D	0.875	0.824
2D+3D	**0.888**	**0.838**

Table 5: Ablation of window size.

window size	MedM
3×3	0.876
5×5	**0.888**
7×7	0.881

Table 6: Ablation of sliding stride.

stride	MedM
1	**0.888**
2	0.885
3	0.886

Ablation on sliding stride. Table 6 shows the comparison of model performance under various stride settings. We evaluate the models at the optimal window size of 5x5 to assess the impact of different strides. The results indicate that the performance differences across strides are subtle, demonstrating the robustness of our approach to changes of stride.

4 Conclusion

We propose a cross-dimension self-supervised learning strategy (CDSSL-P3D) aiming to perform jointly pre-training of 2D and 3D data in medical images. The introduced strategy is not confined to specific network architectures, which can be applied for CNNs and Transformers. We conduct experiments with CNN and Transformer on 13 downstream tasks and compare with a series of advanced SSL methods. Extensive evaluation results amply substantiate the effectiveness of CDSSL-P3D.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, P., Xu, S., Harmon, S.A., Turkbey, E.B., Sanford, T.H., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., Patella, F., Carrafiello, G., Turkbey, B.T., Wood, B.J.: Ct images in covid-19 [data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.GQRY-NC81 (2020)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 ieee. In: CVF International Conference on Computer Vision (ICCV). vol. 3 (2021)
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in neural information processing systems **33**, 12546–12558 (2020)
5. Chellapilla, K., Puri, S., Simard, P.: High performance convolutional neural networks for document processing. In: Tenth international workshop on frontiers in handwriting recognition. Suvisoft (2006)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)

7. Goncharov, M., Soboleva, V., Kurmukov, A., Pisov, M., Belyaev, M.: vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–614. Springer (2023)
8. Haghighi, F., Taher, M.R.H., Zhou, Z., Gotway, M.B., Liang, J.: Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. IEEE transactions on medical imaging **40**(10), 2857–2868 (2021)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
11. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
12. Jin, L., Yang, J., Kuang, K., Ni, B., Gao, Y., Sun, Y., Gao, P., Ma, W., Tan, M., Kang, H., et al.: Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. EBioMedicine **62** (2020)
13. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
14. Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S.: A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence **2**(4), e190138 (2020)
15. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
16. Nguyen, D.M., Nguyen, H., Mai, T.T., Cao, T., Nguyen, B.T., Ho, N., Swoboda, P., Albarqouni, S., Xie, P., Sonntag, D.: Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14426–14435 (2023)
17. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
18. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020)
19. Tsai, Y.M., Luszczek, P., Kurzak, J., Dongarra, J.: Performance-portable autotuning of opencl kernels for convolutional layers of deep neural networks. In: 2016 2nd Workshop on Machine Learning in HPC Environments (MLHPC). pp. 9–18. IEEE (2016)

20. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
22. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
23. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)
24. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
25. Ye, Y., Zhang, J., Chen, Z., Xia, Y.: Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 545–555. Springer (2022)
26. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: Pcrlv2: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. arXiv preprint arXiv:2301.00772 (2023)