Anatomy-based Self-supervised Pre-training for Scale-robust Hierarchical Representations in Chest X-rays*

Surong Chu[1 [0009−0008−4211−7334]], Yan Qiang[2,1 [0000−0001−6231−3721]], Guohua Ji[1 [0009−0004−4573−9543]], Xueting Ren[1 [0000−0002−5424−1868]], Lijing Zhang[1 [0000−0002−7190−0829]], Baoping Jia[4], Yangyang Wei[5], Juanjuan Zhao[1,3 [0000−0001−8134−9076]], and Shuo Li[6,7 [0000−0002−5184−3230]]

[1] College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Taiyuan, China
[2] School of Software, North University of China, Taiyuan, China
[3] Schoool of Software, Taiyuan University of Technology, Taiyuan, China
{zhaojuanjuan}@tyut.edu.cn
[4] Shanxi Cardiovascular Hospital, Taiyuan, China
[5] First Hospital of Shanxi Medical University, Taiyuan, China
[6] the Department of Computer and Data Science, Case Western Reserve University, Cleveland, USA
[7] the Department of Biomedical Engineering, Case Western Reserve University, Cleveland, USA

Abstract. In self-supervised pre-training, learning consistent and hierarchical representations that capture relationships among anatomical semantics holds promise for enhancing the performance and interpretability of downstream tasks. However, the representations learned by existing methods are vulnerable to scale variations, which manifests as inconsistency on some scales and misjudgments of hierarchy. Therefore, we propose a scale-robust anatomical representation learning framework with self-supervision, which incorporates contrastive learning with our newly proposed pretext tasks: location-scale prediction(LSP) and decomposition prediction(DP). Our method addresses the vulnerability from three aspects: 1) It uses multi-scale patches as inputs to embrace diverse anatomical semantics in pre-training. 2) LSP promotes consistency at multi-scales by enhancing the model's sensitivity to scale and resolving representation conflicts caused by multi-scale inputs. 3) DP eliminates hierarchy misjudgments by producing hierarchical representations for anatomies and their constituent parts that better balance the similarity and discriminability. Evaluations across six chest X-ray datasets demonstrate that the representations learned by our method are consistent and hierarchical at multi-scales and have great transferring ability to various downstream tasks. The code is publicly available at https://github.com/SurongChu/SRHRS.

Keywords: Self-supervised learning · consistent · hierarchical representations · anatomical semantics · pre-training.

* The corresponding author is Juanjuan Zhao.

1 Introduction

For medical images with fixed anatomical structures, such as Chest X-Rays (CXRs), learning anatomical representations through Self-Supervised Pre-training (SSP) holds promise to enhance the performance and interpretability of downstream tasks by facilitating the transfer of anatomical knowledge [25, 10, 7]. Unlike traditional representations disassociated with anatomical semantics [18, 28, 26, 24], anatomical representations distinguish themselves by reflecting relationships between anatomical structures. Although the definitions of anatomical representations have evolved from consistent to hierarchical, they invariably face challenges posed by multi-scale anatomical semantics, as shown in Fig.1.

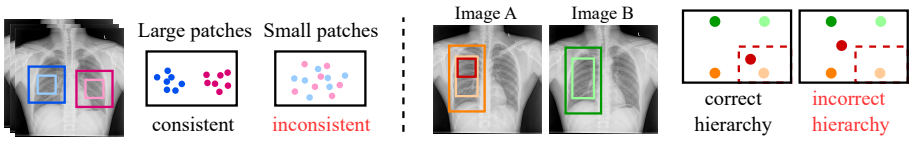

Fig. 1. Challenges to consistent and hierarchical representations caused by multi-scale anatomical semantics.

Early studies [10, 7, 27, 5] believed that anatomical representations should be consistent, where identical anatomical semantics cluster together in the representation space. However, due to fixed-scale inputs during pre-training, these representations are vulnerable to scale variations, resulting in inconsistent representations at some scales, as shown in Fig.1(a).

Recent studies [12, 22] argued that anatomical representations should be hierarchical to reflect the inherent hierarchies among anatomical semantics, such as lungs (which can be divided into left and right lungs, each further subdivided into lobes), ribs, and vessels. Hierarchy, grounded in consistency, necessitates representations to be consistent at multi-scales and requires the representations of anatomical structures show appropriate similarity if one structure is a part of another. Existing methods [12, 22] learned hierarchical representations through coarse-to-fine staged training and bidirectional agreement prediction between anatomical semantics and their constituent parts. However, despite these efforts, the inconsistencies as shown in Fig.1(a) still exist due to the limited diversity of anatomical semantics in their staged training. Furthermore, they encounter a new challenge: misjudgments of hierarchy with similar semantics at multi-scales, as shown in Fig.1(b). This occurs because their excessive agreement constraints on patches and their constituent parts overly encourage similarity but compromise discriminability of representations.

To address the challenges illustrated in Fig.1, we propose two new pretext tasks: Location-Scale Prediction (LSP) and Decomposition Prediction (DP). They are based on the following observations: despite individual variations across

images, anatomies exhibit stable global positions, scales, and invariant relative positions. LSP promotes consistency across various scales by predicting the stable global locations and scales of randomly selected anatomies in CXRs. Although straightforward, LSP fulfills its role by mitigating representation conflicts caused by diverse inputs and enhancing scale sensitivity. DP, on the other hand, leverages the invariant relative position among anatomies to address misjudgments of hierarchy. Specifically, it performs identical decompositions on input patches and their corresponding feature maps and encourages agreement on the decomposed parts of the same locations, thereby better balancing the similarity and discriminability of representations between large anatomical structures and their constituent parts. By integrating these two tasks with Contrastive Learning (CL), we have developed a Scale-Robust Hierarchical Representation learning framework with Self-supervision(SRHRS). It streamlines the staged training process into an end-to-end workflow and embraces a wider range of diverse inputs.

Evaluations across six CXR datasets demonstrate that the representations learned by our SRHRS are consistent and hierarchical at multi-scales and have great transferring ability. Ablation experiments show the interactions among SRHRS's components and the impact of diverse inputs and training manners. Besides, we designed a new experiment, "Finding Parent", which provides quantitative metrics for evaluating the hierarchy of representations.

2 Method

Our SRHRS is built on CL, taking a teacher-student network [6] with multi-scale patches as inputs, see Fig.2(a). A local patch x, cropped from image I with location p and scale s, transformed by $\mathcal{T}$ to get two views x' and x''; and then fed into the teacher (E_{θ_T}) and student encoder (E_{θ_S}) to obtain feature maps $F_T = E_{\theta_T}(x')$, $F_S = E_{\theta_S}(x'')$. These feature maps go through the teacher (G_{ζ_T}) and student projector (G_{ζ_S}) to obtain the projected vectors $z_T = G_{\zeta_T}(F_T)$, $z_S = G_{\zeta_S}(F_S)$. CL learns discriminative features invariant to transformations by minimizing the distance ($dist(\cdot)$) between the projected z_T and z_S with loss $\mathcal{L}_{CL} = dist(z_T, z_S)$. However, simply adopting multi-scale inputs can lead to representation conflicts, thus necessitating the integration of LSP.

2.1 Location-scale prediction(LSP)

To learn consistent representations at multi-scales, LSP predicts the position p and scale s of local patch x. A prediction head H_{η_L} is added after the projector G_{ζ_S} in the student branch (see Fig.2(b)). The LSP loss is the distance between the predicted values $(\hat{p}, \hat{s})$ and the ground truth (p, s), $\mathcal{L}_{LSP} = dist((\hat{p}, \hat{s}), (p, s))$. Location prediction enhances consistency by leveraging the location stability of anatomies, while scale prediction aids in differentiating semantics at the same locations, as CXR anatomical semantics are sensitive to scale variations. Their combination effectively resolves representation conflicts arising from multi-scale inputs, enabling the acquisition of consistent representations at various scales. In

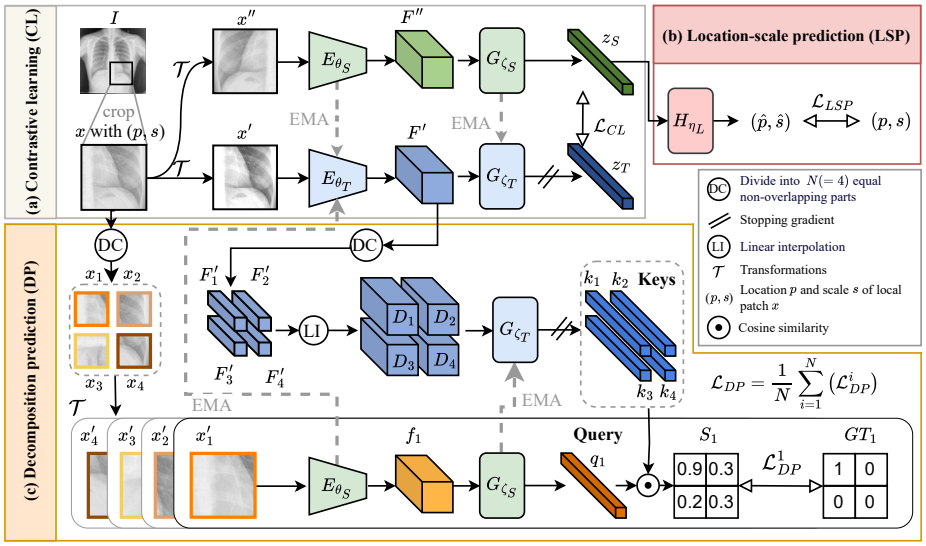

Fig. 2. The architecture of SRHRS, which comprises three tasks: CL, LSP, and DP. CL is a dual-branch contrastive learning framework that takes multi-scale patches as inputs. LSP mitigates representation conflicts and enhances consistency by predicting the locations and scales of input patches. DP fosters hierarchy by strengthening the coherence between corresponding decomposed parts in both feature and image space.

experiments, scale s is selected from set $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and location p is selected from set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, both with a tolerance of ± 0.05 to enhance resilience against individual divergence [3, 17, 11, 16].

2.2 Decomposition prediction (DP)

To address the misjudgments of hierarchy, DP performs identical decompositions in both feature and image space to learn hierarchical representations (see Fig.2(c)): **In feature space**, F' is decomposed into N equal non-overlapping sub-feature maps $\{F'_1, F'_2, \ldots, F'_N\}$, which are then linearly interpolated to produce feature maps $\{D_1, D_2, \ldots, D_N\}$ with the same size as F'. Each D_i $(i = 1, 2, \ldots, N)$ is further projected by G_{ζ_T} to generate a feature vector k_i. Collectively, these N feature vectors form the key set $K = \{k_1, k_2, \ldots, k_N\}$. Similarly, **in image space**, x is decomposed into N parts $\{p_1, p_2, \ldots, p_N\}$. Each p_i, transformed by $\mathcal{T}$, passes through encoder E_{θ_S} and projector G_{ζ_S}, ultimately outputs query vector q_i. The cosine similarities between q_i and all keys are recorded in vector S_i. The ground truth for S_i is the one-hot vector GT_i highlighting k_i because the corresponding image regions of q_i and k_i are largely overlapped. The prediction loss of q_i is the distance between S_i and GT_i, $\mathcal{L}^i_{DP} = dist(S_i, GT_i)$. The DP loss for patch x is the average of the losses across N queries: $\mathcal{L}_{DP} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}^i_{DP}$. In practice, N is set to 4. When $N = 4$, the decomposition only

comprises two basic divisions: one horizontal and one vertical. DP imposes local agreement constraints between a patch and its constituent parts to prevent over-similar representations, thereby preserving discriminability. These representations, which balance similarity and discriminability, can effectively eliminate misjudgments of hierarchy.

2.3 Overall Structure

As shown in Fig.2, our SRHRS consists of three tasks: CL, LSP and DP. The total loss of the SRHRS is the weighted sum of the three tasks' losses: $\mathcal{L} = \mathcal{L}_{CL} + \lambda_{LSP}\mathcal{L}_{LSP} + \lambda_{DP}\mathcal{L}_{DP}$, where L_{CL}, L_{LSP}, and L_{DP} denote the loss of CL, LSP, and DP; λ_{LSP} and λ_{DP} denote the weights of L_{LSP} and L_{DP}, respectively. All these tasks share a dual-branch teacher-student network. The student branch updates via gradient backpropagation, while the teacher updates via exponential moving average (EMA) from the student. After pre-training, only the teacher's encoder is transferred to downstream tasks.

3 Experiments and Results

In this section, we introduce the experiment details and results. The results include: analyses on consistency and hierarchy of representations; evaluation of transferring ability; and ablation studies on the effects of SRHRS's components, diverse inputs and training manners.

3.1 Pre-training settings

The pre-training utilized 67K images from the official training set of Chest X-ray14 [23], while all the remaining images were used for downstream task TDC. Images from a single patient belong exclusively to one set to prevent data leakage. This principle has also been applied to the dataset division for all downstream tasks. The encoders E_{θ_S} and E_{θ_T} take ResNet-50 [9] as backbones, while G_{ζ_S}, G_{ζ_T} and H_{η_L} are composed of two-layer multilayer perceptron (MLP) networks. The parameters of E_{θ_T} and G_{ζ_T} are updated by E_{θ_S} and G_{ζ_S} using EMA with a decay rate of 0.99. The weights in the total loss function were set to the optimal combination where $\lambda_{LSP} = 2$, $\lambda_{DP} = 1$. The $dist(\cdot)$s in $\mathcal{L}_{CL}$, $\mathcal{L}_{LSP}$, $\mathcal{L}_{DP}$ take cross entropy loss. The transformation set $\mathcal{T}$ includes random flipping, rotation, color jittering, and Gaussian blur, followed by resizing to 224×224. Our models are implemented in PyTorch on two NVIDIA A5000 GPUs with 24 GB memory for each, optimized by Adam with a learning rate of 0.0001 and a weight decay of 1e-6. The pre-training takes 300 epochs with early stopping.

3.2 Downstream task settings

We evaluated our method on six downstream tasks, including: (1) a multi-label thorax disease classification (TDC) on a subset of Chest X-ray14; (2) a multi-class pneumoconiosis stages classification (PSC) on dataset SXPCO [5]; (3) a

multi-class COVID-19 classification (CVC) on dataset COVQU [4, 19]; (4) a binary pediatric pneumonia classification (PPC) on dataset GZCP [14]; (5) three independent binary anatomical structure segmentation (ASS) on dataset NIH-Mon [13, 20, 2]; (6) a binary pneumothorax segmentation (PTS) on dataset SIIM-ACR [1]. We follow the official dataset split if provided; if not, we split the dataset into training, validation, and test sets at a 6:2:2 ratio. In TDC, labels are represented by 14-bit binary vectors, with all zeros signifying "no findings".

For a comprehensive evaluation, we compared our approach with training from scratch (Scratch), the fully-supervised pre-trained model on ImageNet (ImageNet), and a wide range of self-supervised methods, including: (1) classic contrastive learning methods: Moco v2 [8] and BYOL [6]; (2) advanced methods for consistent representations: TransVW [7] and GVSL [10]; (3) advanced methods for hierarchical representations: Adam v2 [22]. To ensure a fair comparison, all methods were pre-trained with the same dataset and backbone as SRHRS. Only the pre-trained encoders were used for downstream tasks. A two-layer MLP is added as classification head, a decoder is added as segmentation head to form a U-Net [21], both of which were randomly initialized. Input patches were randomly cropped with a scale of [0.7, 1] and performed the same transformations $\mathcal{T}$. Performance was evaluated using mean Area Under the Curve (AUC) for classification and mean Dice coefficient (DSC) for segmentation. We ran each task 10 times and reported the average.

3.3 Representation analysis

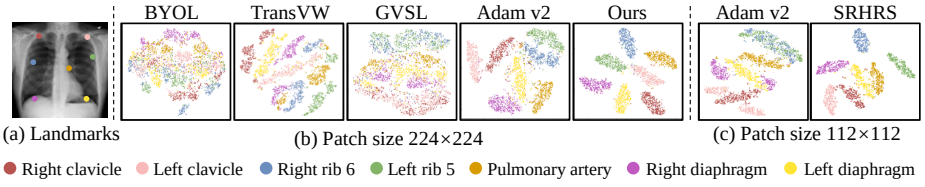

Fig. 3. Representations of patches centered on 7 manually annotated anatomical landmarks are visualized by t-SNE.

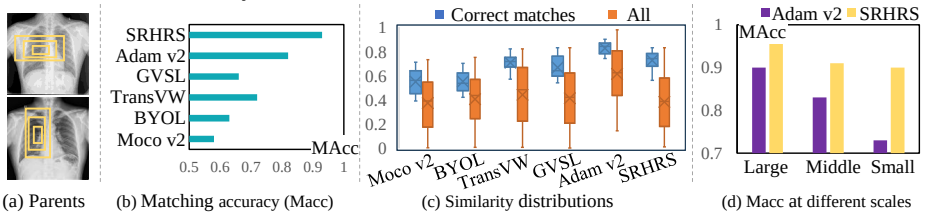

Fig. 4. The experiment "finding parent" is designed to evaluate the hierarchy of representations.

Consistency of representations We compared the five pre-training methods against ours to evaluate the consistency of representations. Firstly, 500 images in the Chest X-ray14's test set were randomly selected and annotated 7 anatomical landmarks by a professional physician (see Fig.3(a)). Subsequently, patches centered on these landmarks were cropped and fed into the pre-trained models to extract representations. Then these representations were visualized by t-SNE [15]. As shown in Fig.3(b), with patch size 224×224, the representations learned by Adam v2 and our SRHRS keep identical anatomical semantics clustered. When the patch size is reduced to 112×112 (see Fig.3(c)), Adam v2 tends to confuse anatomical structures with similar textures (e.g. left rib 5 and right rib 6) and over-dividing those with diverse appearances (e.g. left clavicle and right diaphragm).In contrast, SRHRS excels at distinguishing different anatomical structures. This demonstrates that SRHRS learns consistent representations for anatomical semantics at various scales.

Hierarchy of representations We designed an experiment "finding parent" to evaluate the hierarchy of representations. Firstly, 1000 CXRs were randomly selected from Chest X-ray14's test set and divided into 250 batches. On each image, a group of nested patches were cropped as "parents" (see Fig.4 (a)), with sizes of $0.6H \times 0.3H$, $0.4H \times 0.2H$, and $0.2H \times 0.1H$, where H is the equal length and width of original CXRs. Next, parents were divided equally along its longer edge into child patches. Subsequently, all these patches were input into pre-trained models to extract representations. Finally, each child patch tried to find its parent by selecting the one with the highest cosine similarity to its representation among all the candidates in a batch.

As shown in Fig.4(b), our SRHRS reaches the highest matching accuracy (MAcc), outperforming Adam v2 by over 10%. The similarity distributions of the correct matches and all candidate child-parent pairs are shown in Fig.4(c). SRHRS's similarity distribution for correct matches is lower than Adam v2's, while the distribution for all candidate pairs is more dispersed. This indicates SRHRS's representations create a large disparity between matched and mismatched pairs, and avoid excessive similarity between parent and child patches. When comparing MAcc on patches of different scales (see Fig.4(d)), SRHRS outperforms Adam v2 across all scales, with a significant lead on small patches. In summary, SRHRS's representations are hierarchical at various scales and avoid degrading its discriminative power.

3.4 Transferring learning

We fine-tuned the pre-trained models on six downstream tasks using varying amounts of labeled data to demonstrate their transferring ability (see Tab.1). SRHRS topped all six tasks, not only with full training data but also with limited labeled data. This indicates that SRHRS's consistent and hierarchical representations empower its transferring ability and hold great potential in addressing annotation scarcity.

Table 1. The fine-tuning evaluations on downstream tasks using full training data and 30% training data.

Methods	full training data						30% training data					
	TDC AUC%	PSC AUC%	CVC AUC%	PPC AUC%	ASS DSC%	PTS DSC%	TDC AUC%	PSC AUC%	CVC AUC%	PPC AUC%	ASS DSC%	PTS DSC%
Scratch	73.83	80.20	95.62	94.63	89.53	64.12	53.21	56.40	67.54	66.07	50.42	32.37
ImageNet	75.84	82.03	96.32	94.77	90.25	67.94	56.99	64.17	70.77	72.67	61.22	40.61
Moco v2	74.05	81.72	94.82	94.72	90.35	65.47	54.67	62.18	68.23	70.43	58.29	34.25
BYOL	76.10	81.45	96.33	94.64	89.88	66.24	56.48	63.40	69.19	68.59	60.34	35.83
TransVW	77.93	85.61	97.03	96.68	90.23	67.42	62.72	69.68	77.81	78.12	65.37	45.29
GVSL	76.39	85.07	96.48	95.32	89.38	66.35	57.25	66.34	72.47	74.39	62.08	46.10
Adam v2	78.23	84.78	96.92	96.32	91.04	69.02	64.27	70.82	79.24	78.93	68.41	47.32
SRHRS	**82.42**	**89.43**	**98.32**	**97.85**	**92.77**	**73.35**	**68.57**	**73.16**	**82.93**	**83.28**	**69.28**	**50.25**

3.5 Ablation study

Effects of the components We use the fine-tuning results to evaluate the effect of each SRHRS's components. As shown in Tab.2, CL fails to outperform BYOL in Tab.1, suggesting that multi-scale inputs alone cannot enhance performance due to severe representation conflicts. LSP plays a pivotal role in mitigating these conflicts and learning consistent representations. Furthermore, DP can only fulfill its role effectively when combined with LSP, indicating that consistency is the cornerstone of hierarchy.

Effects of diverse inputs and training manners In Tab.3, we compared patches cropped at fixed (F) or random (R) locations from three or five different scales to show the effect of diverse inputs; and compared end-to-end (E) training with staged (S) one to show the effect of training manners. From bottom to top, we gradually degraded SRHRS (the fourth row) to a simulated staged training approach like [22] (the first row) by reducing the diversity of input patches and flexibility of training manners. To evaluate the effects of these strategies, we utilized the representations to perform the "finding parent" task and reported the MAccs, as this metric directly reflects the hierarchy and consistency of representations. The results indicate that our SRHRS, with diverse inputs and end-to-end training, yields improved representations.

4 Conclusions

We proposed SRHRS, a new SSP framework to address the vulnerability of anatomical representations arising from multi-scale semantics, manifesting as inconsistency at some scales and misjudgments of hierarchy. With our newly proposed pretext tasks LSP and DP, SRHRS successfully addresses this issue by embracing diverse inputs, enhancing scale sensitivity, and balancing the similarity and discriminability of hierarchical representations. Experiments demonstrate that our representations exhibit robust consistency and hierarchy at multi-scales and possess powerful transferring ability to various application scenarios.

Table 2. Fine-tuning evaluation results demonstrate the effects of SRHRS's components.

CL	LSP	DP	TDC AUC%	PSC AUC%	CVC AUC%	PPC AUC%	ASS DSC%	PTS DSC%
√			73.72	80.75	94.53	94.34	87.68	65.39
√	√		80.23	86.26	97.13	96.37	91.24	71.15
√		√	74.84	81.14	95.75	95.07	88.49	66.84
√	√	√	82.42	89.43	98.32	97.85	92.77	73.35

Table 3. Effect of input **Loc**ations (**F**ixed vs. **R**andom) and Scales and **Train**ing manners (**S**taged vs. **E**nd-to-End) on representations

Loc	Scales	Train	MAcc
F	[0.25,0.5,1]	S	0.84
R	[0.25,0.5,1]	S	0.87
R	[0.25,0.5,1]	E	0.88
R	[0.2,0.4,0.6,0.8,1]	E	0.93

Furthermore, SRHRS holds promise to learn consistent and hierarchical representations for other structured medical images, such as CT and MRI.

Acknowledgments. This study was funded in part by the National Natural Science Foundation of China under Grant 62376183, Grant 62476190 and Grant U21A20469, in part by the special fund for Science and Technology Innovation Teams of Shanxi Province under Grant 202304051001009, in part by the Central Government's Guiding Foundation for Local Science and Technology Development under Grant YDZJSX2022C004.

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

References

1. Anna, Z., Carol, W., George, S., Julia, E., Mikhail, F., Mohannad, H., ParasLakhani, Phil, C., Shunxing, B.: Siim-acr pneumothorax segmentation (2019), https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation
2. Brioso, R.C., Pedrosa, J., Mendonça, A.M., Campilho, A.: Semi-supervised multi-structure segmentation in chest x-ray imaging. In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). pp. 814–820. IEEE (2023)
3. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
4. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access **8**, 132665–132676 (2020)
5. Chu, S., Ren, X., Ji, G., Zhao, J., Shi, J., Wei, Y., Pei, B., Qiang, Y.: Learning consistent semantic representation for chest x-ray via anatomical localization in self-supervised pre-training. IEEE Journal of Biomedical and Health Informatics pp. 1–13 (2024). https://doi.org/10.1109/JBHI.2024.3505303
6. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

7. 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)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
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. He, Y., Yang, G., Ge, R., Chen, Y., Coatrieux, J.L., Wang, B., Li, S.: Geometric visual similarity learning in 3d medical image self-supervised pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9538–9547 (2023)
11. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018)
12. Hosseinzadeh Taher, M.R., Gotway, M.B., Liang, J.: Towards foundation models learned from anatomy in medical imaging via self-supervision. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. pp. 94–104. Springer (2023)
13. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
14. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
16. Noothout, J.M., De Vos, B.D., Wolterink, J.M., Postma, E.M., Smeets, P.A., Takx, R.A., Leiner, T., Viergever, M.A., Išgum, I.: Deep learning-based regression and classification for automatic landmark localization in medical images. IEEE transactions on medical imaging **39**(12), 4011–4022 (2020)
17. Pan, Z., Zhuang, B., Liu, J., He, H., Cai, J.: Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/cvf international conference on computer vision. pp. 377–386 (2021)
18. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
19. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
20. Rajaraman, S., Folio, L.R., Dimperio, J., Alderson, P.O., Antani, S.K.: Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations. Diagnostics **11**(4), 616 (2021)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

22. Taher, M.R.H., Gotway, M.B., Liang, J.: Representing part-whole hierarchies in foundation models by learning localizability composability and decomposability from anatomy via self supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11269–11281 (2024)
23. 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)
24. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3024–3033 (2021)
25. Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S.: A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis **69**, 101985 (2021)
26. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Transactions on Medical Imaging **41**(10), 2658–2669 (2022)
27. Yu, K., Sun, L., Chen, J., Reynolds, M., Chaudhary, T., Batmanghelich, K.: Drasclr: A self-supervised framework of learning disease-related and anatomy-specific representation for 3d lung ct images. Medical Image Analysis **92**, 103062 (2024)
28. Zhou, H.Y., Lu, C., Chen, C., Yang, S., Yu, Y.: A unified visual information preservation framework for self-supervised pre-training in medical image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8020–8035 (2023)