Learning Foundation Models from Multi-Organ Medical Images by Capturing Consistency and Diversity of Anatomical Structures

Mohammad Reza Hosseinzadeh Taher[1,2⋆], Junpyo Hong[2], Ravi Soni[2], and Gopal Avinash[2]

[1] Arizona State University, Tempe AZ 85281, USA
mhossei2@asu.edu
[2] GE HealthCare, San Ramon CA 94583, USA
{junpyo.hong,ravi.soni,gopal.avinash}@gehealthcare.com

Abstract. Medical images span a wide range of imaging protocols and anatomical regions, exhibiting two fundamental properties: *inter-organ diversity*–where different organs exhibit distinct structural patterns (e.g., hand vs. chest)–and *intra-organ consistency*–where each organ retains a coherent structure with subtle variations across patient (e.g., left vs. right hand). While existing foundation models typically focus on a single organ or combine organs across heterogeneous modalities–often failing to jointly capture both properties–we envision that a model purposefully built on these fundamental properties would yield representations with greater generalizability, robustness, and interpretability. To this end, we introduce a general-purpose and scalable framework for learning foundation models from diverse organs within a given imaging modality. We call our framework **Coda**, as it is explicitly designed to jointly capture both the **co**nsistency and **d**iversity of **a**natomical structures, encoding high-level semantic relationships across distinct organs and fine-grained anatomical details within each organ. Our experiments in zero-shot, few-shot transfer, and full-transfer settings show that Coda, pretrained on 23 diverse organs, learns semantically rich representations that not only yield strong *inter-organ* and *intra-organ* discrimination capabilities but also offer superior generalizability and robustness on diverse tasks.

Keywords: Foundation models · Multi-organ learning.

1 Introduction

Foundation models like BERT [4] and GPT-4 [23] have led to major breakthroughs in natural language processing, and their success has, in turn, revolutionized the development of vision-language models. A key factor behind their success is their ability to capture the underlying structures (foundation) of the English language, including syntactic and semantic relationships [20]. However,

⋆ Work done during a co-op at GE HealthCare.

foundation models recently developed for medical imaging have yet to achieve comparable impact, largely due to their limited capacity to grasp the foundation of medical images–human anatomy–and to capture the underlying anatomical structures inherent in medical images [12].

Medical images are acquired across a wide range of imaging protocols and anatomical regions and exhibit two fundamental properties: *inter-organ diversity*–where each organ presents distinct structural patterns (e.g., hand vs. chest)–and *intra-organ consistency*–where each organ retains a coherent structure with subtle variations across patients (e.g., left hand vs. right hand). A key limitation of existing foundation models lies in their inability to jointly capture both properties in their learned representations. Models trained on a single organ within a specific modality (e.g., chest X-rays) typically capture intra-organ consistency but lack anatomical diversity, limiting their transferability [26,6]. Conversely, models trained across multiple organs from heterogeneous modalities (e.g., radiographs and CT scans) account for inter-organ diversity but encounter optimization difficulties due to substantial cross-modality discrepancies [28].

To address this limitation, we introduce **Coda**[1], a novel framework explicitly designed to jointly capture both the **co**nsistency and **d**iversity of **a**natomical structures in its learned representations. As illustrated in Fig. 1, Coda serves as a general-purpose and scalable framework for building foundation models from diverse organs within a given imaging modality. Its learned representations are not only *semantically rich*–encoding both high-level relationships across diverse organs and fine-grained anatomical details within each organ–but also *robust* and *transferable* across tasks. Owing to widespread clinical use and anatomical breadth of radiography, we train Coda on a diverse dataset comprising radiographs from **23 distinct organs**. Through extensive evaluations on a variety of downstream tasks, including zero-shot (Fig. 2), full-transfer (Fig. 3), and few-shot (Tab. 1) learning settings, we show that Coda consistently outperforms large-scale fully supervised and self-supervised medical baselines.

Coda fundamentally differs from existing self-supervised learning (SSL) methods [3,14,9,7,1,8,10,27,16,2,11], which disregard semantic organ correlations across images in their learning objectives, by explicitly modeling inter-organ relations across a wide range of images. Coda also sets itself apart from existing supervised and self-supervised multi-organ learning approaches [21,22], which either depend on costly expert annotations or fail to capture the inherent anatomical properties of organs, including inter-organ diversity and intra-organ consistency. In summary, we make the following contributions: **(1)** A novel framework that simultaneously captures inter-organ semantic relationships and fine-grained intra-organ anatomical variations across patients; **(2)** A set of zero-shot analyses demonstrating Coda's capability to model anatomical structures inherent properties, preserving both inter-organ diversity and intra-organ consistency in the learned embedding space; and **(3)** A comprehensive set of experiments showcasing Coda's enhanced generalizability and robustness in few-shot and full-transfer learning settings compared to large-scale fully/self-supervised medical models.

[1] In music, a "coda" is a concluding section that resolves earlier motifs.

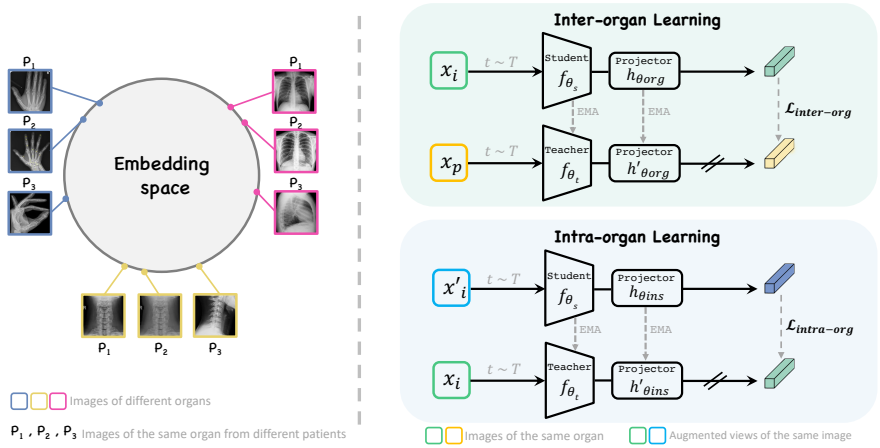

Fig. 1. (Left) Medical images exhibit two key properties: inter-organ diversity, where each organ presents distinct structural patterns (e.g., hand, chest, spine), and intra-organ consistency, where an organ maintains structural resemblance while displaying subtle variations across patients (e.g., left hand vs. right hand). We propose that a deep model capable of capturing these inherent anatomical properties will effectively learn the underlying anatomical structures in medical images, resulting in representations with enhanced generalizability and robustness. To this end, we introduce a novel framework that learns a semantically rich embedding space, effectively distinguishing different organs (denoted by different colored boxes) while also capturing intra-organ variations. Specifically, for a given organ, patients with similar anatomical appearances (e.g., p_1 and p_2) have closer embeddings, while those with distinct anatomical variations (e.g., p3) are mapped farther apart. **(Right)** Coda framework comprises two key branches: (1) inter-organ learning, which captures high-level features to differentiate organ classes by maximizing the agreement between embeddings of the same organ; and (2) intra-organ learning, which encodes fine-grained features to differentiate instances of the same organ by enforcing alignment between embeddings of different views of the same instance. Student and teacher networks are shared among two branches.

2 Method

Our Coda aims to learn semantically rich and robust visual representations by capturing intrinsic anatomical knowledge from multi-organ medical images. As illustrated in Fig. 1, Coda develops a comprehensive understanding of organ structures by modeling semantic organ relationships and capturing fine-grained organ details through two key components:

(1) Inter-organ learning aims to learn high-level discriminative representations that enable the clustering of images of the same organ while distinguishing images of different organs (e.g., hand, chest, spine) in the embedding space. The inter-organ branch consists of the student (f_{θ_s}) and teacher (f_{θ_t}) encoders, as well as two projector heads $h_{\theta_{org}}$ and $h'_{\theta_{org}}$. The parameters of the student f_{θ_s} and the projector head $h_{\theta_{org}}$ are optimized via back-propagation, whereas the

teacher network f_{θ_t} and head $h'_{\theta_{org}}$ are updated via an exponential moving average (EMA) of the f_{θ_s} and $h_{\theta_{org}}$, respectively. Given an input batch of images, we apply data augmentation twice to create two augmented copies, which together form a multi-viewed batch. The first augmented copy of the batch is processed by the student and projector head $h_{\theta_{org}}$, while the second augmented copy is processed by the teacher and projector head $h'_{\theta_{org}}$. For any sample x_i within the multi-viewed batch, the objective of the inter-organ branch is to maximize the agreement between its embeddings and those of the samples $x_p \in P_{x_i}$ that share the same organ as x_i, while minimizing the similarity between its embeddings and those of the samples $x_n \in N_{x_i}$ from different organs. To this end, we employ a contrastive loss function [18] tailored for multi-organ datasets:

$$\mathcal{L}_{inter-org} = -\frac{1}{|P_{x_i}|} \sum_{x_p \in P_{x_i}} \log \frac{\exp\left(z_i \cdot z_p/\tau\right)}{\sum_{x_n \in N_{x_i}} \exp\left(z_i \cdot z_n/\tau\right)} \tag{1}$$

where τ is a temperature parameter that controls distribution sharpness.

(2) Intra-organ learning aims to learn fine-grained discriminative representations that enable identifying subtle differences within instances of the same organ (e.g. left vs. right hand). The intra-organ branch consists of the student (f_{θ_s}) and teacher (f_{θ_t}) encoders, which are shared with the inter-organ branch, along with two projector heads $h_{\theta_{ins}}$ and $h'_{\theta_{ins}}$. Given an input sample x_i, we first extract a set N of multi-scale crops from it. We then apply data augmentation to x_i and pass it through the teacher network and projector $h'_{\theta_{ins}}$ to obtain its embedding $z_i^t = h'_{\theta_{ins}}(f_{\theta_t}(x_i))$. Next, we augment the multi-scale crops in N, and pass them through the student network and projector $h_{\theta_{ins}}$ to obtain their embeddings $Z_i^s = \{z_i^s = h_{\theta_{ins}}(f_{\theta_s}(x'_i)) \mid x'_i \in N\}$. The objective of the intra-organ branch is to maximize the agreement between the softmax normalized embeddings [1] of the input sample and its augmented views. To achieve this, we employ a cross-entropy loss:

$$\mathcal{L}_{intra-org} = -\frac{1}{|Z_i^s|} \sum_{z_i^s \in Z_i^s} -z_i^t \, log \, z_i^s \tag{2}$$

Overall training scheme. To enable end-to-end representation learning, the inter-organ and intra-organ learning objectives are combined into a total loss $\mathcal{L} = \mathcal{L}_{inter-org} + \mathcal{L}_{intra-org}$. Through our unified training scheme, Coda learns an embedding space where semantically similar organs have similar embeddings while preserving subtle variations within instances of the same organ, resulting in more powerful representations for diverse tasks.

3 Experiments and Results

We adopt the base version of ConvNeXt (ConvNeXt-B) [19] as the backbone for both student and teacher networks. Projection heads consist of 3-layer MLPs with a hidden dimension of 2048, l_2 normalization, and a weight-normalized fully

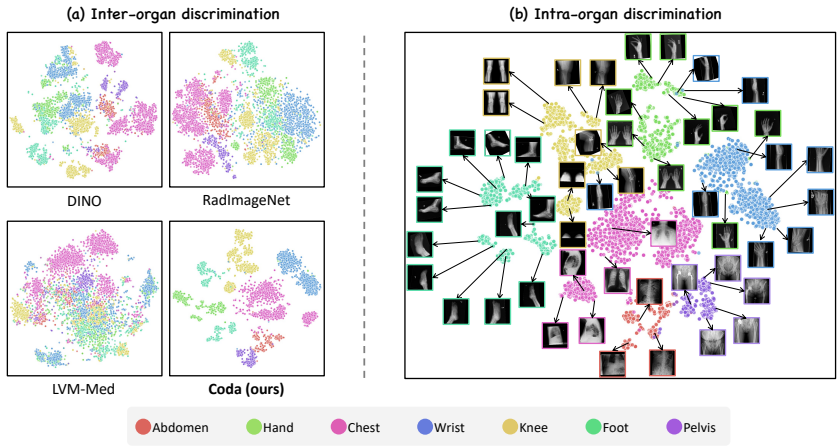

Fig. 2. **(a)** t-SNE visualization of organ embeddings (unseen in Coda's training) in zero-shot setting (without fine-tuning). Unlike existing multi-organ models, Coda produces semantics-rich embeddings with strong inter-organ discrimination, forming well-separated clusters for different organs (denoted by distinct colors). **(b)** A detailed analysis of Coda's embeddings reveals that within each organ class, distinct islands emerge, corresponding to anatomical variations of the same organ. This demonstrates Coda's intra-organ discrimination capabilities. For example, knee images (yellow) form separate clusters based on imaging views (sunrise, lateral, and anteroposterior); chest images (pink) split into frontal and lateral views; and hand images (green) group by different hand poses. Additionally, anatomically related regions, such as the pelvis and abdomen, form closer clusters, reflecting their inherent structural similarities.

connected layer with an output dimension of 65,536. Coda is trained on a diverse dataset of radiographs covering 23 distinct organs, including the abdomen, ankle, calcaneus, chest, clavicle, elbow, facial bones, femur, finger, foot, forearm, hand, hip, humerus, knee, neck, patella, pelvis, shoulder, skull, spine, thumb, and wrist. We train Coda for 800 epochs using the AdamW optimizer, batch size of 64, a cosine learning rate scheduler with a base learning rate of $1.25e-4$, and input resolution 224×224. Data augmentations T include color jittering, Gaussian blur, rotation, and random cropping with $N = 12$ crops of size 96×96 and a scale range of [0.4, 0.7]. Once trained, the *teacher* model is transferred to downstream tasks. In the following, Coda is rigorously compared with both fully-supervised and self-supervised baselines trained on large-scale datasets, demonstrating superior performance in zero-shot anatomy understanding, few-shot transfer, and full-transfer settings across 6 common yet challenging tasks on 5 public datasets, covering various diseases and body parts.

(1) Coda encodes semantics-rich representation, excelling in anatomy understanding in zero-shot settings.

Experimental Setup: To evaluate Coda's anatomical understanding capabilities, we analyze its representations in a zero-shot setting (without fine-tuning) against

state-of-the-art supervised and self-supervised multi-organ medical models, including RadImageNet [21] and LVM-Med [22]. Specifically, we use images *unseen* during Coda's pretraining, containing seven randomly selected organs (abdomen, knee, hand, foot, wrist, pelvis, and chest), extract their embeddings using Coda and each baseline model, and visualize them in a 2D space using t-SNE plots.

Result and Analysis: As shown in Fig. 2-a, both the fully-supervised RadImageNet and self-supervised LVM-Med models produce mixed embeddings, where different organs are not well-separated. By contrast, Coda generates highly separable features, resulting in distinct clusters that clearly distinguish different organs. This underscores Coda's ability to learn discriminative features that enable strong inter-organ discrimination. An in-depth analysis of the t-SNE plot for Coda in Fig. 2-a reveals distinct islands within each organ class. We delve deeper into this phenomenon by inspecting specific images within each island. Fig. 2-b shows that within different islands, images of the same organ with varying appearances appear, highlighting the intra-organ discrimination capabilities of Coda's representations. For example, knee images (yellow points) form separate islands based on sunrise, lateral, and anteroposterior views. Similarly, chest images (pink points) are grouped by frontal and lateral views. Also, hand images (green points) are separated into groups containing two hands, left hands, and right hands. Moreover, closely related anatomical regions, such as the pelvis and abdomen, form clusters closer together compared to other organs, reflecting their structural similarities. In conclusion, these results demonstrate that Coda not only captures high-level discriminative features by modeling the semantic relationships between organs, but also effectively extracts fine-grained anatomical details within individual images, enabling distinct separation of images at both the class and instance levels.

(2) Coda offers generalizable representations for a variety of tasks, demonstrating superiority in full-transfer settings.

Experimental Setup: To demonstrate the generalizability of the representations learned by our framework, we compare the transfer performance of Coda against competitive publicly available fully-supervised and SSL baselines trained on large-scale datasets, including DINO [1], RadImageNet [21], LVM-Med [22], and CheSS [3], which are trained on 1.2M, 1.3M, 1.35M, and 4.8M images, respectively. We fully fine-tune all models on 6 diverse downstream tasks, including tuberculosis [17], lung nodule [25], and pneumothorax [29] classification, as well as clavicle [17], heart [17], and femur [5] segmentation. Adhering to standard transfer learning protocols [15,13], pretrained models are adapted for classification tasks by appending a task-specific fully connected layer to pretrained models. For segmentation tasks, a U-Net architecture [24] is employed, with the encoder initialized using pretrained models and the decoder randomly initialized. To ensure fair comparisons, we utilize the official model of each baseline, which is meticulously optimized, and run each model 10 times under the same experimental setup in each task. We also report performance when training the downstream models from random initialization.

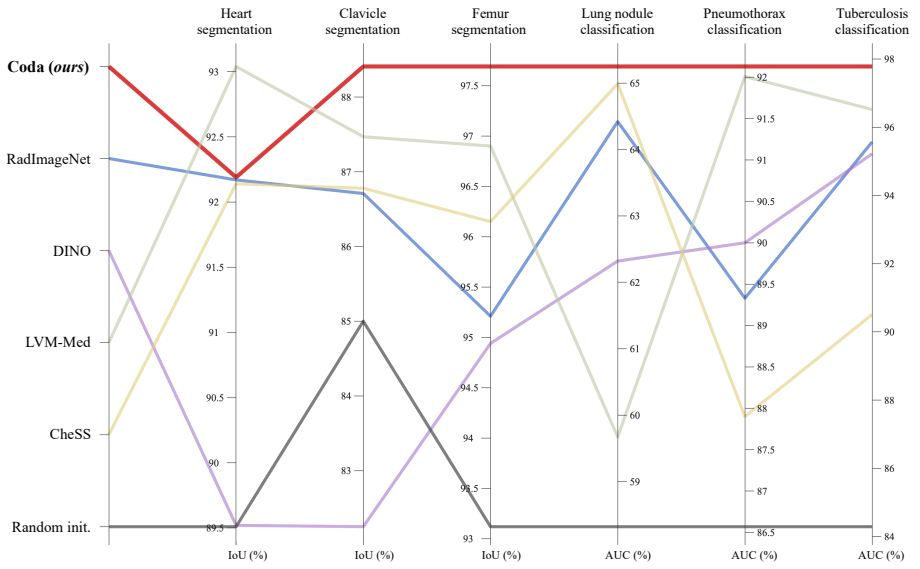

Fig. 3. Coda offers generalizable representations for a variety of tasks, demonstrating superior transfer performance across diverse classification/segmentation tasks compared with state-of-the-art baselines, including large-scale fully supervised and self-supervised multi-organ medical models.

Result and Analysis: As shown in Fig. 3, Coda consistently exhibits superior performance compared to fully-supervised and self-supervised baselines across tasks. Specifically, Coda outperforms the supervised RadImageNet model across all tasks, despite being pretrained using SSL without relying on human expert annotations, whereas RadImageNet depends on such annotations. Furthermore, Coda surpasses self-supervised DINO and CheSS baselines in all tasks. Additionally, Coda surpasses LVM-Med, which was trained on 1.3 million medical images from 16 body organs, achieving superior performance in tuberculosis, nodule, and pneumothorax classification, as well as clavicle and femur segmentation. This highlights the effectiveness of Coda in learning diverse and discriminative representations from a variety of organs by capturing their semantic relationships in the embedding space, leading to improved transferability across different tasks, diseases, and organs.

(3) Coda provides robust representations for limited data regimes, achieving superior performance in few-shot transfer settings.

Experimental Setup: We assess the robustness of Coda's representations in limited data regimes, a critical requirement for medical imaging applications with scarce annotated data. To this end, we conduct few-shot transfer learning experiments on tuberculosis classification, clavicle segmentation, and heart segmentation. Specifically, we fine-tune Coda using varying portions of labeled data

Table 1. Coda achieves remarkably superior performance in few-shot transfer for classification and segmentation tasks, highlighting the significance of our anatomy learning strategy in enhancing the robustness of the learned representations. Also, with only 10%, 10%, and 20% of the training data, Coda achieves 93%, 91%, and 93% of its full-data performance in tuberculosis classification, heart segmentation, and clavicle segmentation, demonstrating its strong data efficiency. Δ show performance boosts achieved by Coda compared with the baseline in each task/data portion.

Method	Tuberculosis classification (AUC%)				Clavicle segmentation (IoU%)				Heart segmentation (IoU%)			
	1% (4-shot)	10%	20%	50%	1% (1-shot)	10%	20%	50%	1% (1-shot)	10%	20%	50%
LVM-Med	59.50	86.82	88.78	94.27	50.54	53.49	69.10	79.88	36.42	48.04	48.15	73.97
Coda (*ours*)	**81.70**	**90.72**	**93.09**	**96.14**	**56.26**	**76.96**	**82.50**	**85.97**	**67.05**	**83.75**	**88.42**	**91.13**
Δ	↑22.2	↑3.9	↑4.3	↑1.9	↑5.7	↑23.5	↑13.4	↑6.1	↑30.6	↑35.7	↑40.3	↑17.2

(1%, 10%, 20%, and 50%) and compare its performance with LVM-Med, which outperformed other state-of-the-art supervised and self-supervised models in full-transfer settings (see Fig. 3).

Result and Analysis: As shown in Tab. 1, Coda consistently outperforms LVM-Med across all tasks and data fractions, achieving average performance gains of 8%, 30%, and 12% in tuberculosis classification, heart segmentation, and clavicle segmentation, respectively. Notably, in tuberculosis classification, when fine-tuned with only 4 training samples (1% of the data), Coda surpasses LVM-Med by a substantial margin of 22.2%. Similarly, in heart and clavicle segmentation, using just 1 training sample (1% of the data), Coda exceeds LVM-Med by 30.6% and 5.7%, respectively. Furthermore, with only 10%, 10%, and 20% of the training data, Coda achieves 93%, 91%, and 93% of its full-data performance in tuberculosis classification, heart segmentation, and clavicle segmentation, respectively, demonstrating its strong data efficiency. These results highlight the effectiveness of our anatomy-guided learning strategy in enhancing the robustness of the learned representations, enabling Coda to generalize effectively in low-data settings and ultimately reducing annotation costs.

(4) Ablation: impact of learning objectives. We evaluate the impact of each learning branch in Coda by comparing the performance of each branch individually with the performance when both branches are integrated across two downstream tasks. As seen in Tab. 2, the integration of $\mathcal{L}_{inter-org}$ and $\mathcal{L}_{intra-org}$ results in superior performance compared to using either objective individually. This suggests that the unified learning approach in Coda effectively captures both high-level organ discrimination and fine-grained organ details, leading to more comprehensive representations across downstream tasks.

4 Conclusion

This work introduces Coda, a general-purpose and scalable framework for learning foundation models from diverse organs within a given imaging modality. By explicitly modeling both the consistency and diversity of anatomical structures,

Table 2. Ablation study on each learning objective of Coda: The integration of $\mathcal{L}_{inter-org}$ and $\mathcal{L}_{intra-org}$ within our framework yields more diverse and comprehensive representations, resulting in superior performance over either objective individually.

	$\mathcal{L}_{inter-org}$	$\mathcal{L}_{intra-org}$	Tuberculosis classification (AUC%)	Clavicle segmentation (Dice%)
Objective	✓		95.74±0.96	84.85±0.94
		✓	97.17±0.74	85.23±0.54
	✓	✓	**97.78±0.77**	**85.59±0.60**

Coda captures semantically rich representations that encode high-level relationships across organs and preserve fine-grained intra-organ variations. Through extensive evaluations in zero-shot, few-shot, and full-transfer settings, Coda demonstrates strong inter-organ and intra-organ discrimination, as well as superior generalizability and robustness across medical tasks. In future work, we aim to extend Coda to additional imaging modalities to further broaden its applicability.

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

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020)
3. Cho, K., Kim, K.D., Nam, Y., Jeong, J., Kim, J., Choi, C., Lee, S., Lee, J.S., Woo, S., Hong, G.S., Seo, J.B., Kim, N.: CheSS: Chest x-ray pre-trained model via self-supervised contrastive learning. Journal of Digital Imaging (2023)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2018), https://arxiv.org/abs/1810.04805
5. Gut, D.: X-ray images of the hip joints (2021). https://doi.org/10.17632/zm6bxzhmfz.1, https://data.mendeley.com/datasets/zm6bxzhmfz/1
6. Haghighi, F., Gotway, M.B., Liang, J.: Learning anatomy-disease entangled representation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4129–4141 (2025)
7. Haghighi, F., Hosseinzadeh Taher, M.R., Gotway, M.B., Liang, J.: Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? Medical Image Analysis **94**, 103086 (2024)
8. Haghighi, F., Hosseinzadeh Taher, M.R., Zhou, Z., Gotway, M.B., Liang, J.: Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 137–147 (2020)

9. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20824–20834 (2022)
10. 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). https://doi.org/10.1109/TMI.2021.3060634
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022)
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. Hosseinzadeh Taher, M.R., Haghighi, F., Feng, R., Gotway, M.B., Liang, J.: A systematic benchmarking analysis of transfer learning for medical image analysis. In: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health. pp. 3–13 (2021)
14. Hosseinzadeh Taher, M.R., Haghighi, F., Gotway, M.B., Liang, J.: Caid: Context-aware instance discrimination for self-supervised learning in medical imaging. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. vol. 172, pp. 535–551 (2022)
15. Hosseinzadeh Taher, M.R., Haghighi, F., Gotway, M.B., Liang, J.: Large-scale benchmarking and boosting transfer learning for medical image analysis. Medical image analysis **102**, 103487 (2025)
16. Hosseinzadeh Taher, M.R., Hong, J., Soni, R., Avinash, G.: Moss: Learning from multiple organs via self-supervision. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2025)
17. 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) (2014)
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
19. Liu, Z., Mao, H., Wu, C.Y., et al.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
20. Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., Levy, O.: Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences **117**(48), 30046–30054 (2020)
21. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., et al.: Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
22. Nguyen, D.M.H., Nguyen, H., Diep, N.T., Pham, T.N., Cao, T., Nguyen, B.T., et al.: Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching (2023)
23. OpenAI: Gpt-4 technical report (2023), https://arxiv.org/abs/2303.08774

24. 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)
25. Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule. American Journal of Roentgenology **174**(1), 71–74 (2000). https://doi.org/10.2214/ajr.174.1.1740071, https://doi.org/10.2214/ajr.174.1.1740071, pMID: 10628457
26. 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 (CVPR). pp. 11269–11281 (2024)
27. Taher, M.R.H., Ikuta, M., Soni, R.: Curriculum self-supervised learning for 3d ct cardiac image segmentation. In: ML4H@NeurIPS (2023), https://api.semanticscholar.org/CorpusID:267760581
28. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11114–11124 (2024)
29. Zawacki, A., Wu, C., Shih, G., Elliott, J., Fomitchev, M., Hussain, M., Lakhani, P., Culliton, P., Bao, S.: Siim-acr pneumothorax segmentation 2019 (2019), https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/