Supplementary Materials for Representing Part-Whole Hierarchies in Foundation Models by Learning Localizability, Composability, and Decomposability from Anatomy via Self-Supervision

Mohammad Reza Hosseinzadeh Taher[1] Michael B. Gotway[2] Jianming Liang[1]
[1]Arizona State University [2]Mayo Clinic
{mhossei2,jianming.liang}@asu.edu Gotway.Michael@mayo.edu

A. Overview

- **Goal.** The primary goal of this paper is to highlight the importance of constructing a hierarchy of embeddings that can be autodidactically learned from anatomy and that can enhance the generalizability and robustness of representation learning.
- **Hypothesis.** We hypothesize that if deep neural networks can comprehend images akin to human perception—parsing them into part-whole hierarchies [12–14]—their learned features would exhibit increased generalizability, robustness, and interpretability.
- **Challenge.** While deep neural networks excel in learning multi-level feature spaces, their limitation often lies in the absence of explicit coding for part-whole hierarchies, hindering a nuanced understanding of hierarchical relationships among objects and their constituent parts [12, 21].
- **Solution.** We propose a novel self-supervised learning framework–called Adam-v2–that, *without* requiring anatomy labeling, explicitly incorporates part-whole hierarchies into its learning objectives through three key branches—localizability, composability, and decomposability—in order to preserve a semantic balance of anatomical diversity and harmony in its learned embedding space (§2).
- **Contributions.** In addition to higher generalizability and transferability of our Adam-v2's learned representations (Fig. 7 and Tab. 2), our Adam-v2 proves to be an effective few-shot learner, making it a potent pretraining model for segmentation tasks with a scarcity of annotations (Tab. 1 and 4). Furthermore, we present a comprehensive set of quantitative and qualitative feature analyses that offer new perspectives for assessing anatomy understanding from various viewpoints (§4.1).
- **Relation to GLOM.** Hinton recently introduced the idea of "GLOM" [12], aiming to signify the importance of explicitly presenting part-whole hierarchies in a neural network. Inspired by the conceptual idea beneath GLOM, we propose Adam-v2 which is *fundamentally* different from GLOM in several key aspects. Firstly, GLOM is an *imaginary system* without practical experimentation, whereas the Adam-v2 is a *functioning system* rigorously evaluated across 10 tasks in diverse settings. Secondly, GLOM is an idea for developing *an ideal architecture* for constructing hierarchical representations. However, our Adam-v2 offers *a functional framework* that takes a step towards achieving the overarching goal shared with GLOM—interpreting images as part-whole hierarchies akin to human vision systems—through a simple yet effective *learning strategy* that does not rely on labeled data. A recent line of research works has attempted to implement GLOM, such as [8, 28]. However, our work diverges from this line of research in that Adam–v2 encodes the semantics of part-whole hierarchies into the embedding space through training with three explicit objectives: localizability, composability, and decomposability.

B. Additional Results

B.1. Few-shot Transfer in Fundus Imaging Tasks

To underscore the effectiveness of our SSL framework in learning robust representations for few-shot segmentation tasks, we replicate the experiments reported in Tab. 1 within few-shot transfer settings for fundus applications. To do so, we fine-tune our Adam–v2 model and baseline models, all of which were pretrained on fundus images from the EyePACS dataset, using a limited number of labeled samples (3-shot, 6-shot, and 12-shot) from the DRIVE and Drishti-GS datasets. As seen in Tab. 4, Adam–v2 exhibits superior few-shot transfer performance when compared to SSL methods in retinal vessel segmentation and optic disk segmentation tasks on the DRIVE and Drishti-GS datasets, respectively. Notably, when compared to the runner-up baseline, Adam–v2 achieves improvements of 5.92%, 4.25%,

Method	DRIVE (Dice%)			
	3-shot	6-shot	12-shot	Full-shot
DINO [5]	68.96	71.49	72.73	78.36
DenseCL [33]	67.99	70.65	72.83	78.36
DiRA [11]	68.13	70.80	72.88	78.52
Adam–v2 (Ours)	**74.87**	**75.74**	**76.27**	**79.91**
Δ	+5.92	+4.25	+3.39	+1.39

Method	Drishti-GS (Dice%)			
	3-shot	6-shot	12-shot	Full-shot
DINO [5]	85.29	92.36	93.95	96.44
DenseCL [33]	85.42	92.51	94.10	96.60
DiRA [11]	86.42	92.81	94.14	96.76
Adam–v2 (Ours)	**94.00**	**94.97**	**96.20**	**97.02**
Δ	+7.57	+2.16	+2.07	+0.26

Table 4. Adam–v2 demonstrates superior performance in few-shot transfer within the fundus modality, surpassing SOTA SSL methods by a large margin in the retinal vessel and optic disk segmentation tasks on the DRIVE and Drishti-GS datasets, respectively. Remarkably, with only 3 shots, Adam–v2 achieves 94% and 97% of its full training data performance in retinal vessel and optic disk segmentation tasks, respectively. Δ shows Adam–v2's performance boosts compared with second-best method (underlined). All methods are pretrained on Eye-PACS dataset.

and 3.39% in 3-shot, 6-shot, and 12-shot scenarios for the retinal vessel segmentation task, and 7.57%, 2.16%, and 2.07% for the optic disk segmentation task. With only 3 shots, Adam–v2 achieves 94% and 97% of its full training data performance in retinal vessel and optic disk segmentation tasks, respectively.

B.2. Weakly-supervised Disease Localization

We explore the effectiveness of our Adam–v2 in localizing chest pathology in a weakly supervised setting. To do so, we follow [32, 34] and leverage the ChestX-ray14 dataset, which comprises 787 cases annotated with bounding boxes for eight thorax diseases: Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, and Pneumothorax. During the training phase, we initialize target models with our Adam–v2 and other baselines pretrained weights and fine-tune them using only image-level disease labels. In the testing phase, we employ Grad-CAM [9] to visualize image regions responsible for the model predictions, specifically identifying the diseased regions. As seen in Fig. 9, Adam–v2 localizes diseases more accurately compared to other baselines. Notably, the heatmaps generated by our Adam–v2 exhibit a higher degree of concentration around disease regions in comparison to other baselines, with its attention maps displaying a more pronounced overlap with the ground truth across all diseases. This generation of more interpretable activation maps not only highlights the Adam–v2's potential for precise disease localization but also shows its potential for clinical utility in post-hoc interpretation by radiologists.

B.3. Anatomy Matching

To further illustrate Adam–v2's capability in anatomy understanding, we examine Adam–v2's representations for anatomical landmark matching in a zero-shot setting. To do so, from a given query image, we randomly select $N_q = 13$ anatomical landmark points and extract a patch of size 96^2 centered at each anatomical landmark point. These patches are resized to 224^2 and passed through Adam–v2's pretrained backbone to generate query embeddings $E_q = \{E_q^i\}_{i=1}^{N_q}$. Then, for an *unlabeled* key image, we extract N_k patches by sliding a window of size 96^2 with a stride of 16. After resizing these key patches to 224^2, their embeddings $E_k = \{E_k^j\}_{j=1}^{N_k}$ are obtained using Adam–v2's pretrained backbone. Finally, for each query anatomical landmark embedding in E_q, we compute its ℓ_2-distance with all embeddings in E_k and identify the center of the patch corresponding to the embeddings with the minimum distance as the matched point for the anatomical landmark.

To showcase the robustness of Adam–v2's representations to anatomical variations, we explore three distinct settings involving anatomical point matching across images of the same patient with different diseases, images of different patients, and augmented views of the same image. Fig. 10 depicts the query image annotated with 13 landmark points (red circles) and corresponding matched points (indicated by yellow crosses) across different diseases, patients, and views. To assess the accuracy of matched points, we include the ground truth landmark points provided by human experts for key images (depicted as red circles). As shown in Figure 10, Adam–v2 showcases its potential in precisely identifying similar anatomical landmarks, aligning with our findings in Sec. 4.1 and emphasizing the semantic richness inherent in Adam–v2's representations. These findings underscore an additional emergent property of our Adam–v2, revealing its potential in identifying corresponding landmarks. It is crucial to emphasize that our primary focus in this study is to acquire generalizable and semantically rich representations through our proposed SSL learning strategy. For an in-depth exploration of Adam–v2's potential in landmark detection and image registration—topics beyond the scope of this paper—a more detailed investigation is warranted, which we defer to future work.

C. Downstream Tasks

We extensively evaluate the transferability of Adam–v2 pretrained models across a broad spectrum of 10 downstream tasks on nine publicly available datasets of chest X-ray and fundus modalities. These tasks assess the generalizability

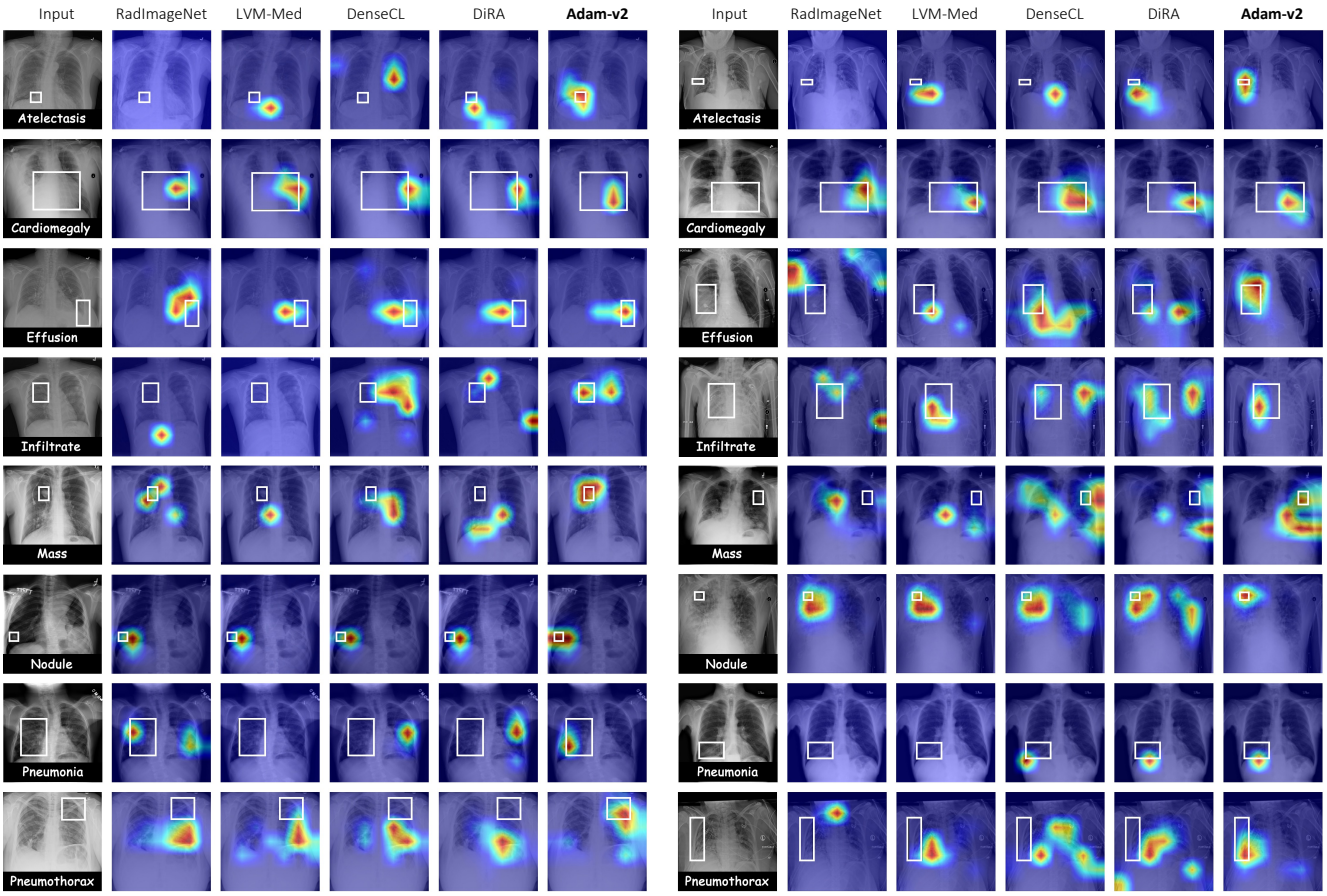

Figure 9. [Better viewed on-line, in color, and zoomed in for details] Visualization of Grad-CAM heatmaps generated by Adam–v2 and baseline methods for eight diseases in ChestX-ray14. Ground truth is marked with white boxes. Adam–v2 yields finer localization outcomes compared to baselines, which either concentrate on broader image regions or miss alignment with the ground truth.

Adam–v2's representations across various applications (binary classification, multi-label classification, organ/lesion segmentation), diseases (tuberculosis, pneumothorax, lung nodules, etc), anatomical structures (heart, clavicle, ribs, vessels, optic disk), and modalities (chest radiography and fundus photography). In the following, we elaborate on the specifics of downstream tasks incorporated in this paper.

Task 1: Clavicle segmentation. This task entails pixel-level segmentation of the left and right clavicles. We use the Japanese Society of Radiological Technology (JSRT) dataset [25, 30], comprising 247 posterior-anterior chest radiographs with associated segmentation masks for clavicles. The dataset was divided into two folds, containing 124 and 123 images, respectively. We adhere to the official patient-wise data split, utilizing fold-1 for training and fold-2 for testing. We use mean Dice score as the evaluation metric for assessing clavicle segmentation performance.

Task 2: Heart segmentation. This task encompasses pixel-level segmentation of the heart. We use JSRT dataset [25, 30] for this task, comprising 247 images with associated segmentation masks for heart. We adhere to the official patient-wise data split, utilizing fold-1 (124 images) for training and fold-2 (123 images) for testing. We use mean Dice score as the evaluation metric for assessing heart segmentation performance.

Task 3: Ribs segmentation. This task entails pixel-level segmentation of individual ribs. We utilize the VinDr-Rib dataset [22], comprising 245 chest radiographs accompanied by segmentation masks for 20 individual anterior and posterior ribs (10 on each side of the lungs). Following the official dataset split, we use 196 images for training and 49 for testing. This task is formulated as a multi-class segmentation problem, and the performance is evaluated using the mean Dice score.

Task 4: Thoracic diseases segmentation. This task involves pixel-level segmentation of thoracic diseases using ChestX-Det dataset [20]. The dataset comprises 3,578 chest X-ray images. Board-certified radiologists have provided

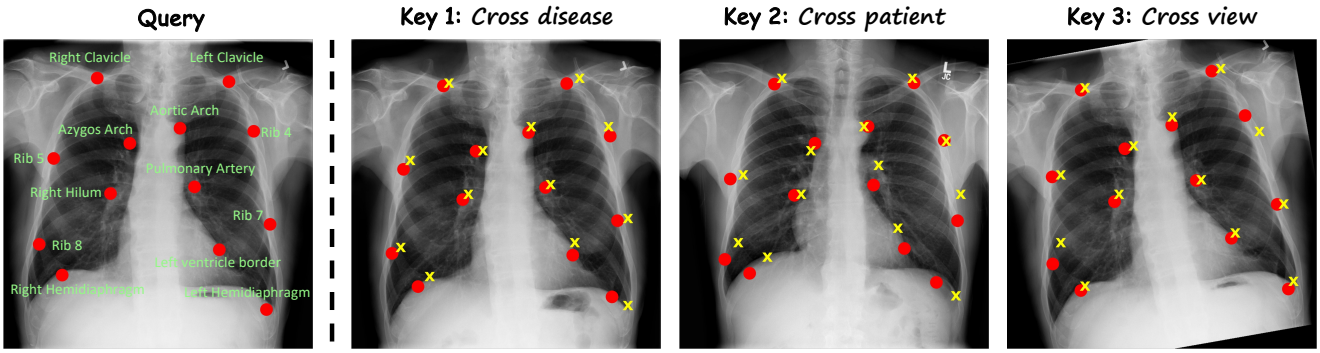

Figure 10. [Better viewed on-line, in color, and zoomed in for details] Adam–v2 shows its potential in identifying similar anatomical landmarks across three distinct settings: anatomical point matching across images of the same patient with different diseases, images of different patients, and augmented views of the same image. The images are from the test set of the ChestX-ray14 dataset. The red circles represent the ground truth for 13 distinct landmark points in query and key images, while the yellow crosses indicate the corresponding matched points identified by our Adam–v2.

segmentation masks for 13 common thoracic conditions, including atelectasis, calcification, cardiomegaly, consolidation, diffuse nodule, effusion, emphysema, fibrosis, fracture, mass, nodule, pleural thickening, and pneumothorax. We adhere to the official dataset split, using 3,025 images for training and 553 images for testing.

Task 5: Pneumothorax segmentation. This task focuses on pixel-level segmentation of pneumothorax disease. We utilize the SIIM-ACR dataset [35], consisting of 10,000 chest radiographs along with segmentation masks for pneumothorax disease, if present in an image. For training and testing, we randomly divide the dataset into 8,000 and 2,000 images, respectively. Segmentation performance is assessed using the mean Dice score.

Task 6: Common thoracic diseases classification. This task involves multi-label classification of five common thoracic diseases. We use VinDR-CXR dataset [23] that provides 18,000 posterior-anterior chest radiographs, along with image-level labels provided by expert radiologists for 6 conditions: lung tumor, pneumonia, tuberculosis, other diseases, COPD, and No finding. Following the official dataset split, we allocate 15,000 images for training and 3,000 for testing, and evaluate classification performance using the mean AUC score.

Task 7: Tuberculosis classification. This task involves detection of tuberculosis disease. We use NIH Shenzhen CXR dataset [17], including 662 chest radiographs, with 326 images categorized as normal and 336 images representing patients with tuberculosis. We randomly split the dataset into training (80%) and testing (20%) sets, and evaluate performance using the AUC metric.

Task 8: Thoracic diseases classification. This task encompasses multi-label classification of fourteen thoracic diseases, employing the ChestX-ray14 dataset [32] curated by the National Institutes of Health Clinical Center, USA. The dataset comprises 112,120 de-identified X-rays from 30,805 unique patients, with labels indicating the absence or presence of 14 thoracic disease categories. We adhere to the official patient-wise split provided by the dataset, allocating 86K images for training and 25K for testing, and assess classification performance using the mean AUC over the 14 diseases.

Task 9: Retinal vessel segmentation. This task encompasses pixel-level segmentation of retinal vessels. We use DRIVE dataset [4], including 40 color fundus images along with expert annotations for retinal vessels. Following the official dataset split, we use 20 images for training and 20 for testing, and evaluate segmentation performance using the mean Dice score.

Task 10: Optic disk segmentation. This task involves pixel-level segmentation of optic disk. We use Drishti-GS dataset [26], encompassing 101 fundus images, divided into 50 training and 51 testing images. Ground truth segmentation masks are provided for optic disk by human experts. We adhere to the official dataset split and assess segmentation performance using the mean Dice score.

D. Implementation Details

D.1. Pretraining Setup

Our SSL framework is architecture-neutral and compatible with any ConvNet and vision transformer backbones. We have trained two Adam–v2 models with ResNet-50 backbone using unlabeled images from the training sets of ChestX-ray14 [32] and EyePACS [7] datasets for chest X-ray and fundus imaging tasks, respectively. Moreover, we have trained Adam–v2 with ViT-S backbone using unlabeled images from the training sets of ChestX-ray14 and

CheXpert [16] datasets. Additionally, to demonstrate the scalability of our framework, we have trained a large-scale Adam–v2 model with ConvNeXt-B backbone using a large corpus of 926K chest X-ray images collected from 13 publicly-available datasets, including ChestX-ray14, CheXpert [16], VinDr-CXR [23], NIH Shenzhen CXR [17], RSNA Pneumonia Detection Challenge [27], MIMIC-CXR [18], PadChest [2], COVID-19 Radiography Database [6], Indiana ChestX-ray [1], Mendeley-V2 [19], COVIDx [31], JSRT [30], and NIH Montgomery [17]. Following [5], the localizability heads $h_{\theta_{LS}}$ and $h_{\theta_{LT}}$ consist of a 3-layer multi-layer perceptron (MLP) with hidden dimension 2048 and output dimension $K = 65536$. The composability (h_{θ_C}) and decomposability (h_{θ_D}) heads are 2-layer MLP with hidden dimension 2048. We use AdamW optimizer, and follow [5] in learning rate scheduler and weight decay settings. To empower the model with hierarchical anatomy learning, we train Adam–v2 in a coarse-to-fine manner, incorporating diverse anatomical structures at various scales. Starting with $m = 0$, where the model is trained on the entire anatomy (whole images), we progressively reduce the scale of anatomical structures by factors based on powers of 2. Specifically, for input images with spatial resolution $(H \times W)$, we randomly sample anatomical structures with resolutions of $(\frac{H}{2^m} \times \frac{W}{2^m})$, where $m \in \{1, 2, ...\}$, and utilize them as inputs to the model during the pretraining process. For learning anatomical structures at each scale, the model is trained with the objective function in Eq. (5). In practice, we assess anatomical structure resolutions across up to 4 scales (i.e., $m \in \{0, .., 3\}$), but our ablation study (see Fig. 8) suggests that up to three levels are sufficient to yield robust representations. During the training, the parameters of the teacher network and localizability head $h_{\theta_{LT}}$ are updated with an exponential moving average on the weights of the student network and $h_{\theta_{LS}}$, respectively; the update rules are $\theta_T \leftarrow \lambda\theta_T + (1-\lambda)\theta_S$ and $\theta_{LT} \leftarrow \lambda\theta_{LT} + (1-\lambda)\theta_{LS}$, where λ follows a cosine schedule from 0.996 to 1 during training [10]. Following [5], we use centering and sharpening for the teacher's outputs to avoid collapsing solutions for localizability learning. In the localizability branch, we extract one global crop of size 224^2 from the input w, along with eight multi-scale crops of size 96^2. The temperature τ_s is set to 0.1, and τ_t follows a linear warm-up as [5]. We train ResNet-50 model from scratch with a batch size 512 distributed across 8 Nvidia V100-32Gb GPUs. We first warm up the localizability branch with a scheme as [5] (200/200/100 epochs for $m = 0, 1, 2$), empowering the model with an initial ability to discriminate different anatomical structures. Subsequently, the composability and decomposability losses are integrated into the training process, and the entire framework is jointly trained (10/90/165 epochs for $m = 0, 1, 2$). We initialize the ConvNeXt-B model with ImageNet-22K pretrained weights, and train it with a batch size 160. We first warm up the localizability branch (70/70/30 epochs for $m = 0, 1, 2$), and then train the model with all three losses (10/50/30 epochs for $m = 0, 1, 2$).

D.2. Downstream Setup

Evaluations. We utilize the pretrained teacher backbone of Adam–v2 (i.e., g_{θ_T}) for zero-shot, few-shot, and full transfer evaluations. We use Adam–v2 with ResNet-50 backbone in zero-shot, few-shot, and full transfer evaluations. We use Adam–v2 model with ConvNeXt backbone for comparison with large-scale medical models in public ChestX-ray14 benchmark. We use Adam–v2 with ViT-S backbone for co-segment visualizations (Fig. 1), and follow the settings of [3] to co-segment the common chest anatomical structures.

Fine-tuning settings. For transfer learning to segmentation tasks, we employ a U-Net architecture [24], initializing the encoder weights with Adam–v2's pretrained backbone. For transfer learning to classification tasks, we take Adam–v2's pretrained backbone and append a fully connected layer to generate the desired classification outputs. Following the standard evaluation protocol [15, 29], we perform end-to-end fine-tuning for all parameters of the target models across all downstream tasks. We strive to optimize each downstream task with the most effective hyperparameters. In classification tasks, we use AdamW optimizer with learning rate $2.5e-4$ decayed by a cosine schedule, weight decay 0.05, $(\beta_1, \beta_2) = (0.9, 0.95)$, and standard data augmentation, encompassing random crop, flip, and rotation. In segmentation tasks, we use Adam optimizer with learning rate $1e-3$ for VinDR-Ribs, DRIVE, and Drishti-GS datasets, and AdamW optimizer with a learning rate $2e-4$ for the rest of the tasks, cosine learning rate decay scheduler, and standard data augmentation, encompassing random crop, brightness contrast, grid and optical distortion, elastic transformation, gamma. Moreover, we employ early-stopping using 10% of the training data as the validation set. We use input size 512^2 for DRIVE and 224^2 for all other tasks. We follow [15] for ChestX-ray14 dataset. Classification and segmentation performances are measured by the AUC (area under the ROC curve), and mean Dice coefficient metrics and IoU (Intersection over Union) metrics, respectively.

E. Acknowledgements

We acknowledge that Zhou et al. first hypothesized, observed, and illustrated the phenomena [36] that emerge from their PEAC's anatomical embedding space via co-segmentation [3], which were coined as "echo-chambers" earlier by Hinton [12], but the "echo-chambers" shown in Fig. 1 (in the main paper) produced by Mohammad Reza Hosseinzadeh Taher are what directly and automat-

ically emerged from our Adam–v2's embeddings (Eve–v2). This research has been supported in part by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and in part by the NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work has utilized the GPUs provided in part by the ASU Research Computing and in part by the Bridges-2 at Pittsburgh Supercomputing Center through allocation BCS190015 and the Anvil at Purdue University through allocation MED220025 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The content of this paper is covered by patents pending.

References

[1] Chest x-rays (indiana university). https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university. 5

[2] Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. 5

[3] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022. 5

[4] A Budai, R Bock, A Maier, J Hornegger, and G Michelson. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013. 4

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2, 5

[6] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. 5

[7] Jorge Cuadros and George Bresnick. Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening. *Diabetes Science and Technology*, 3(3):509–516, 2009. 4

[8] Nicola Garau, Niccolò Bisagno, Zeno Sambugaro, and Nicola Conci. Interpretable part-whole hierarchies and conceptual-semantic relationships in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13689–13698, 2022. 1

[9] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 2

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 5

[11] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. 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)*, pages 20824–20834, 2022. 2

[12] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35(3):413–452, 2023. 1, 5

[13] Geoffrey E. Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cogn. Sci.*, 3:231–250, 1979.

[14] Geoffrey E. Hinton. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46(1):47–75, 1990. 1

[15] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. 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*, pages 3–13, Cham, 2021. Springer International Publishing. 5

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv:1901.07031*, 2019. 5

[17] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6), 2014. 4, 5

[18] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. 5

[19] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification, 2018. 5

[20] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 3

[21] Ramy Mounir, Sujal Vijayaraghavan, and Sudeep Sarkar. STREAMER: Streaming representation learning and event segmentation in a hierarchical manner. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[22] Hoang C. Nguyen, Tung T. Le, Hieu H. Pham, and Ha Q Nguyen. Vindr-ribcxr: A benchmark dataset for automatic

segmentation and labeling of individual ribs on chest x-rays. In *Medical Imaging with Deep Learning*, 2021. 3

[23] Ha Q. Nguyen and et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020. 4, 5

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 5

[25] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174(1):71–74, 2000. 3

[26] Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A. Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 53–56, 2014. 4

[27] Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, Marc Kohli, Mark McDonald, Phil Culliton, and Safwan Halabi- and Tian Xia. Rsna pneumonia detection challenge, 2018. 5

[28] Shuyang Sun, Xiaoyu Yue, Song Bai, and Philip Torr. Visual parser: Representing part-whole hierarchies with transformers, 2022. 1

[29] Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. *arXiv:2309.15358*, 2023. 5

[30] B. van Ginneken, M.B. Stegmann, and M. Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006. 3, 5

[31] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(19549), 2020. 5

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017. 2, 4

[33] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3024–3033, 2021. 2

[34] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3588–3600, 2023. 2

[35] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. 4

[36] Ziyu Zhou, Haozhe Luo, Jiaxuan Pang, Xiaowei Ding, Michael Gotway, and Jianming Liang. Learning anatomically consistent embedding for chest radiography. In *Proceedings of the 34th British Machine Vision Conference (BMVC 2023)*, 2023. 5