—Supplementary Material— Towards Foundation Models Learned from Anatomy in Medical Imaging via Self-Supervision

Mohammad Reza Hosseinzadeh Taher[1], Michael B. Gotway[2], and Jianming Liang[1]

[1] Arizona State University, Tempe, AZ 85281, USA
{mhossei2,jianming.liang}@asu.edu
[2] Mayo Clinic, Scottsdale, AZ 85259, USA
Gotway.Michael@mayo.edu

A The intuition behind our proposed SSL framework

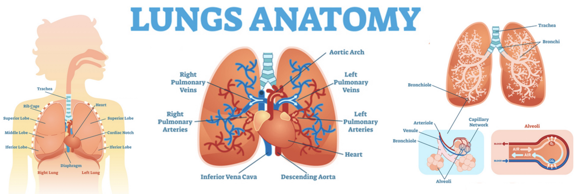

Fig. 6. Human anatomy exhibit natural hierarchies. For example, lung divided into right and left lung, each with lobes. The right lung has three lobes: superior, middle, and inferior; the left lung has two lobes: superior and inferior. The pulmonary arteries, veins, and airways form hierarchical trees. These anatomy hierarchies have inspired us to propose a SSL strategy that captures locality and compositionality of anatomical structures in its embedding space, crucial for anatomy understanding, yet overlooked in existing SSL methods. The image for the lung anatomy available at https://stock.adobe.com/.

B Adam's capability in anatomy understanding

We delve deeper into Adam's capability to generate semantics-rich dense embeddings, where different anatomical structures are associated with different embeddings, and the same anatomical structures have (nearly) identical embeddings at all resolutions and scales. To do so, we employ a dataset comprising 1,000 images along with 4 distinct anatomical landmarks annotated in each image (details in Sec. 3.3). We then extract three patches of different resolutions, denoted as levels 1, 2, and 3, around each landmark location across the images. As a result,

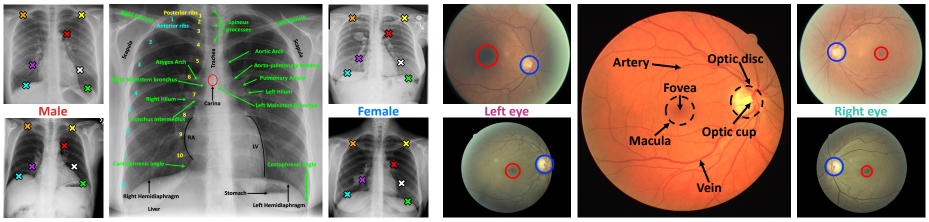

Fig. 7. The anatomical similarity of medical images generated from a particular imaging protocol yields consistent hierarchical anatomical structures, which can be placed at different spatial locations across images due to inter-subject variations. This paper exploits the intrinsic anatomical hierarchies in medical images for SSL, yielding consistent anatomical embeddings without relying on *spatial* correspondence across patients.

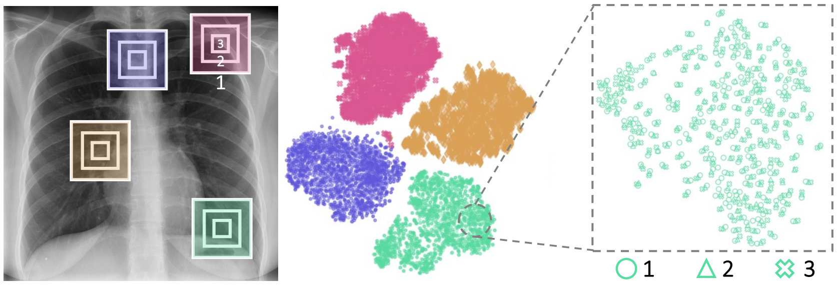

Fig. 8. Adam is capable of generating semantics-rich dense embeddings (Eve), where different anatomical structures are associated with different embeddings, and the same anatomical structures have (nearly) identical embeddings at all resolutions and scales.

instances of each of the four distinct anatomical landmarks represent different anatomical structures. Furthermore, the anatomical structures corresponding to these four landmarks at level 1 exhibit close similarity to their corresponding structures at levels 2 and 3. All anatomical structures in each level are resized to 224×224, and Adam's pretrained model is used to extract their embeddings (i.e. Eve). Finally, tSNE was used to visualize the embeddings. As seen in Fig. 8, the instances of four distinct anatomical landmarks (represented by four different colors) are well-separated from one another, highlighting Adam's capability in distinguishing different anatomical structures. Moreover, the embeddings of the anatomical structures at levels 1, 2, and 3 for each of the four landmarks are close to each other, echoing Adam's ability to provide (almost) identical embeddings for similar anatomical structures across different resolutions.

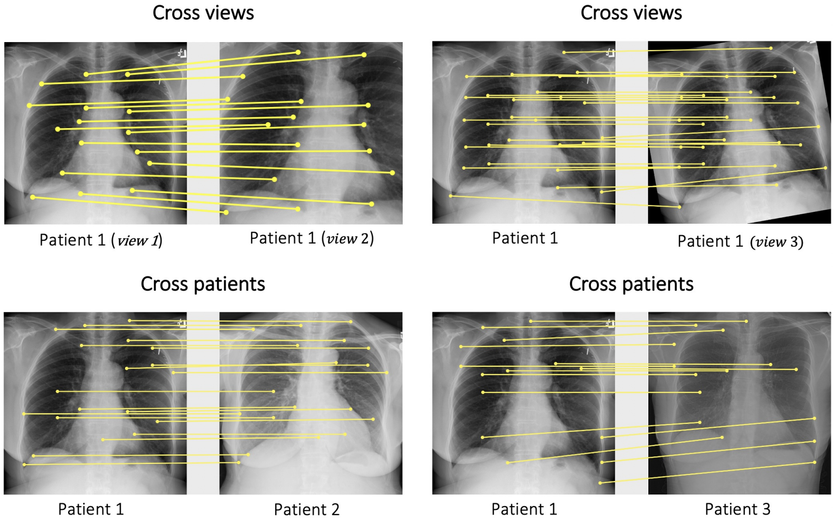

Fig. 9. Visualization of dense correspondence provided by Eve across different views of the same image (first row) and different patients with diversity in intensity distribution and organs' appearance (second row).

C Additional results

C.1 Dense correspondence visualization

To further demonstrate the Eve's accuracy in anatomy understanding, we explore the Eve's robustness to (i) image augmentations and (ii) variations in appearance, intensity, and texture of anatomical structures caused by inter-subject differences or data distribution shifts. To do so, we visualize the dense correspondence between (i) an image and its augmented views produced by cropping and rotation (10 degrees) and (ii) images of different patients with considerable diversity in intensity distribution, texture, and organs' shape. For clarity of figures, we only show some of the high-similarity matches. A match between two feature vectors is represented by a yellow line. Fig. 9 shows Eve is capable of finding similar anatomical patterns across the different views or even across patients. We conclude that Eve provides accurate anatomical representations, mapping semantically similar anatomical structures, regardless of their subtle differences in shape, intensity, and texture, to similar embeddings. Although our method is not designed for this purpose, these results show its potential for landmark detection and image registration applications. It should be noted that our method's primary goal is to provide generalizable models; thus, while our Eve shows some potential for dense visual correspondence, more detailed investigation and comparisons with SOTA methods in this context, such as [31], are required, which we leave to future work.

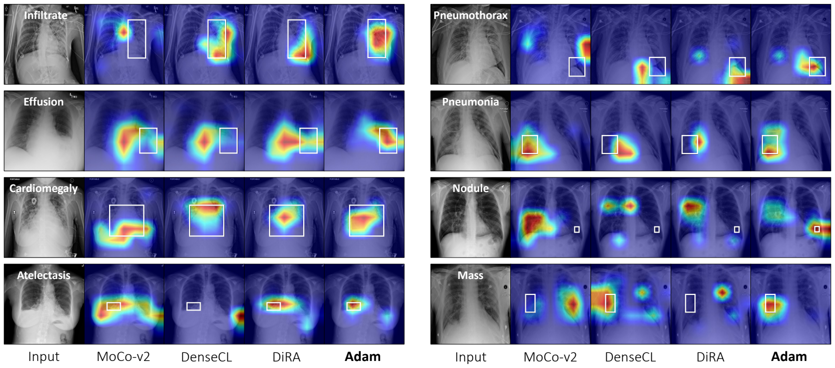

Fig. 10. Visualization of Grad-CAM heatmaps generated by Adam and the bestperforming SSL methods for eight diseases in ChestX-ray14. White boxes indicate ground truth. Adam provides more precise localization results than baselines that focus on larger image regions or fail to overlap with the ground truth.

C.2 GradCAM visualizations for disease localization

We further assess the efficacy of Adam's representations for weakly-supervised disease localization. To do so, we use ChestX-ray14 dataset, which provides bounding box annotations of 8 abnormalities for around 1,000 test images. The images with bounding box annotations are only used during the testing phase to evaluate the localization accuracy. For training, we initialize the downstream model with Adam's pretrained weights and fine-tune it using only image-level disease labels. Then, following [27], we calculate heatmaps using GradCAM to approximate the spatial location of a particular disease. We compare Adam with the best performing SSL methods from each baseline group (i.e. instance-level, patch-level, and pixel-level). Fig. 10 shows examples of GradCAM for Adam and other SSL baselines in eight thoracic diseases, including *Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, Pneumothorax.* As seen, Adam captures the diseased areas more precisely than the baselines. In particular, SSL baselines' attention maps either focus on larger image regions or don't overlap with the ground truth, whereas Adam provides more robust localization results across all diseases. These findings highlight Adam's ability to learn dense representations that are more useful for disease localization.

C.3 Ablation study on pruning threshold

To explore the impact of pruning threshold (γ) of our PP module on the performance of downstream tasks, we have conducted extensive ablation studies on different values of γ. To do so, we pretrain Adam with three pruning thresholds 0.7, 0.8, and 0.9, and transfer the pretrained model with each pruning threshold to three downstream tasks, including SCR-Heart, SIIM-ACR, and ChestX-Det.

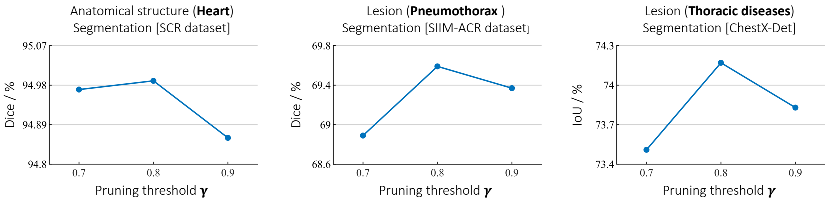

Fig. 11. We conduct ablation study on the impact of pruning threshold on the downstream task performance on three downstream tasks. The best performance achieved with $\gamma = 0.8$ in all applications.

Algorithm 1: Purposive Pruner

```
Input: Anchor embeddings $q$;
       Granularity level $n$;
       Pruning threshold $\gamma$;
       Memory bank MB;
Output: Pruned memory bank $\mathrm{MB}_{\mathrm{pruned}}$
if $n = 0$ then
    $\mathrm{MB}_{\mathrm{pruned}} = \mathrm{MB}$ ;
else
    // remove semantically similar patches to anchor from the memory
       bank
    // sim(x,y) = $\frac{x}{\|x\|_2} \cdot \frac{y}{\|y\|_2}$
    foreach $k_i \in \mathrm{MB}$ do
        if $sim(k_i, q) < \gamma$ then
            $\mathrm{MB}_{\mathrm{pruned}} \leftarrow k_i$ ;
    end
end
```

Fig. 11 depicts the performance of Adam on three downstream tasks under different pruning thresholds. The best performance achieved at $\gamma = 0.8$ in all applications.

D Purposive pruner algorithm

Algorithm 1 presents the details of our purposive pruner (PP) component.

E Datasets and downstream tasks

We pretrain Adam on two publicly available datasets, and thoroughly evaluate the transfer capability of Adam's representations in a wide range of 9 challenging downstream tasks on 8 publicly available datasets in chest X-ray and fundus modalities. In the following, we describe the details of datasets and downstream tasks used in our study.

(1) ChestX-ray14—multi-label classification: ChestX-ray14 dataset provides 112K chest radiographs taken from 30K unique patients, along with 14 thoracic disease labels. Each individual image may have more than one disease label. The downstream task is a multi-label classification in which the models are trained to predict 14 diseases for each image. We use the official patient-wise split released by the dataset, including 86K training images and 25K testing images. We use mean AUC over 14 diseases to evaluate the multi-label classification performance. Moreover, we use the unlabeled training data for pretraining of Adam and other self-supervised baselines.

(2) NIH Shenzhen CXR—binary classification: NIH Shenzhen CXR dataset provides 662 frontal-view chest radiographs, among which 326 images are normal and 336 images are patients with tuberculosis (TB) disease. The downstream task is a binary classification in which the models are trained to detect TB in images. We randomly divide the dataset into a training set (80%) and a test set (20%). We report AUC score to evaluate the classification performance.

(3) VinDR-CXR—multi-label classification: VinDR-CXR dataset provides 18,000 postero-anterior (PA) view chest radiographs that were manually annotated by a total of 17 experienced radiologists for the classification of 5 common thoracic diseases, including pulmonary embolism, lung tumor, pneumonia, tuberculosis, and other diseases. The dataset provides an official split, including a training set of 15,000 scans and a test set of 3,000 scans. We utilize the official split, and report AUC score to evaluate the classification performance.

(4) SIIM-ACR—lesion segmentation: SIIM-ACR dataset provides 10K chest radiographs, including normal cases and cases with pneumothorax disease. For diseased cases, pixel-level segmentation masks are provided. The downstream task is pneumothorax segmentation. We randomly divided the dataset into training (80%) and testing (20%). We use mean Dice score to evaluate segmentation performance.

(5) ChestX-Det—lesion segmentation: ChestX-Det dataset consists of 3,578 images from ChestX-ray14 dataset. This dataset provides segmentation masks for 13 thoracic diseases, including atelectasis, calcification, cardiomegaly, consolidation, diffuse nodule, effusion, emphysema, fibrosis, fracture, mass, nodule, pleural thickening, and pneumothorax. The images are annotated by 3 board-certified radiologists. The downstream task is pixel-wise segmentation of abnormalities in images. We randomly divided the dataset into training (80%) and testing (20%). We use the mean IoU score to evaluate the segmentation performance.

(6) SCR-Heart&Clavicle—organ segmentation: SCR dataset provides 247 posterior-anterior chest radiographs from JSRT database along with segmentation masks for the heart, lungs, and clavicles. The data has been subdivided into two folds with 124 and 123 images. We follow the official split of the dataset, using fold1 for training (124 images) and fold2 for testing (123 images). We use the mean Dice score to evaluate the heart and clavicles segmentation performances.

(7) VinDR-Rib—organ segmentation: VinDR-Rib dataset contains 245 chest radiographs that were obtained from VinDr-CXR dataset and were manually labeled by human experts. The dataset provides segmentation annotations for 20 indivisual ribs. We use the official split released by the dataset, including a training set of 196 images and a validation set of 49 images. We use mean Dice score to evaluate segmentation performance.

(8) EyePACS—self-supervised pretraining: EyePACS dataset consists of 88,702 colour fundus images. Expert annotations for the presence of Diabetic Retinopathy (DR) with a scale of 0–4 were provided for each image. The dataset provides an official split, including 35,126 samples for training and 53,576 samples for testing. We use unlabeled training images for self-supervised pretraining of Adam and other SSL baselines.

(9) DRIVE—organ segmentation: The Digital Retinal Images for Vessel Extraction (DRIVE) dataset includes 40 color fundus images along with expert annotations for retinal vessel segmentation. The set of 40 images was equally divided into 20 images for the training set and 20 images for the testing set. We use the official data split and report the mean Dice score for the segmentation of blood vessels.

F Implementation details

F.1 Pretraining protocol

In our training strategy, we use a standard ResNet-50 as the backbone in accordance with common protocol [28, 16, 10]. Any other sophisticated backbones (i.e., variants of convolutional neural networks or vision transformers) can, however, be leveraged in our proposed training strategy. In this study, we aim to dissect the importance of training strategy in blazing the way for learning generalizable representaitons. As such, we control other confounding factors, including the pretraining data. Consequently, Adam and all self-supervised baseline methods are pretrained on the same pretraining data from ChestX-ray14 and EyePACS datasets. We closely follow the settings of [7] for the training parameters, including the architecture of projection heads (i.e. two-layer MLP), memory bank size (i.e. $K = 65536$), contrastive temperature scaling (i.e. $\tau = 0.2$), and momentum coefficient (0.999). We use even values for n and continue the training process up to $n = 4$, but one can continue the training process with finer data granularity levels. It should be noted that our PP module impose negligible computational cost to the pretraining stage. We use a batch size 256 distributed across 4 Nvidia V100 GPUs with a memory of 32 GB per-card. At each training stage n, we train the model for 200 epochs.

F.2 Fine-tuning protocol

We transfer Adam's pretrained backbone (i.e., f_θ) to the downstream classification tasks by appending a task-specific classification head. For the downstream segmentation tasks, we employ a U-Net network with a ResNet-50 encoder, where the encoder is initialized with the pre-trained backbone. Following the standard protocol [10, 12], we evaluate the generalization of Adam's representations by fine-tuning all the parameters of downstream models. We use input image resolution 224×224 and 512×512 for downstream tasks on chest X-ray and fundus images, respectively. We endeavor to optimize each downstream task with the best-performing hyperparameters as follows. For downstream classification tasks, we use standard data augmentation techniques, including random rotation by $(-7, 7)$ degree, random crop, and random horizontal flip with probability 0.5. We follow [29] in training settings, including AdamW optimizer with weight decay 0.05, $\beta_1, \beta_2 = (0.9, 0.95)$, learning rate $2.5e-4$, and cosine annealing learning rate decay scheduler. For downstream segmentation tasks, we use standard data augmentation techniques, including random gamma, elastic transformation, random brightness contrast, optical distortion, and grid distortion. We use Adam optimizer with learning rate $1e-3$ for VinDR-Ribs and AdamW optimizer with a learning rate $2e-4$ for the rest of the tasks. We use cosine learning rate decay scheduler and early-stopping using 10% of the training data as the validation set. We run each method ten times on each task and report the average, standard deviation, and statistical analysis based on an independent two-sample t-test.

G Acknowledgements

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.