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

Abstract

Humans effortlessly interpret images by parsing them into part-whole hierarchies; deep learning excels in learning multi-level feature spaces, but they often lack explicit coding of part-whole relations, a prominent property of medical imaging. To overcome this limitation, we introduce Adam–v2, a new self-supervised learning framework extending Adam [69] by explicitly incorporating part-whole hierarchies into its learning objectives through three key branches: (1) Localizability, acquiring discriminative representations to distinguish different anatomical patterns; (2) Composability, learning each anatomical structure in a parts-to-whole manner; and (3) Decomposability, comprehending each anatomical structure in a whole-to-parts manner. Experimental results across 10 tasks, compared to 11 baselines in zero-shot, few-shot transfer, and full fine-tuning settings, showcase Adam–v2's superior performance over large-scale medical models and existing SSL methods across diverse downstream tasks. The higher generality and robustness of Adam–v2's representations originate from its explicit construction of hierarchies for distinct anatomical structures from unlabeled medical images. Adam–v2 preserves a semantic balance of anatomical diversity and harmony in its embedding, yielding representations that are both generic and semantically meaningful, yet overlooked in existing SSL methods. All code and pretrained models are available at GitHub.com/JLiangLab/Eden.

1. Introduction

Human perception effortlessly parses visual scenes into *part-whole* hierarchies [39–41]. For instance, when interpreting a chest radiograph, even untrained observers can quickly form a hierarchy by dividing the lower respiratory tract into the left and right lungs, whereas more experienced observers can invoke further sub-hierarchies (see Sec. 1). Deep learning has enabled breakthroughs in learning visual representation at multiple levels. However, the multi-level feature space learned by deep models does not explicitly code part-whole hierarchies with necessary semantic information to indicate hierarchical relationships among wholes and their constituent parts [39, 59].

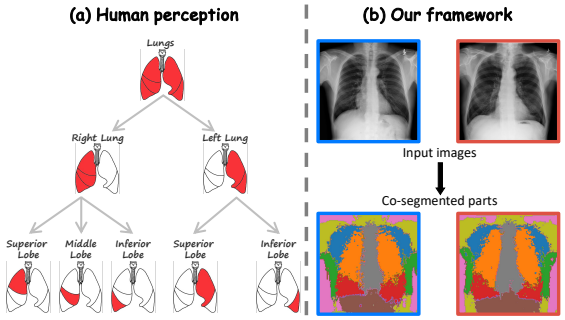

Figure 1. Human perception effortlessly organizes objects into hierarchies to understand their part-whole relationships in images. Taking lungs as an example in (a), even a non-radiologist can form a hierarchy of the right and left lungs, whereas a radiologist can further "see" the lobes in sub-hierarchies. To emulate this ability, we introduce a self-supervised learning framework that *explicitly* learns to encode inherent part-whole hierarchies within medical images into an embedding space, leading to the development of a powerful model (Adam–v2) that is foundational to medical imaging. Adam-v2 can transform each pixel in medical images (e.g., chest radiographs in (b)) into semantically meaningful embeddings (Eve–v2), forming multiple "*echo chambers*" (produced via co-segmentation [1, 97])—different anatomical structures are associated with distinct embeddings, and the same anatomical structures have (nearly) identical embeddings across patients.

To mimic the human ability to understand part-whole hierarchies in images, Hinton, in his *idea* paper [39], introduced an *imaginary* system (*i.e.*, GLOM), aiming to signify the importance of explicitly presenting part-whole hierarchies in a neural network. Inspired by the conceptual idea underlying GLOM, we devise a novel self-supervised learning (SSL) framework, leading to a *functioning* system

that, from medical images, *autodidactically* constructs a hierarchy of embeddings for distinct anatomical structures, semantically balancing anatomical diversity and harmony at each level and conveying parental "whole" at the higher level and filial "parts" at the lower level.

Our framework, as illustrated in Fig. 2, comprises three branches: (1) "localizability", which compels the model to learn a semantically structured embedding space by discriminating between different anatomical structures, (2) "composability", which empowers the model to learn part-whole relations by constructing each anatomical structure through the integration of its constituent parts, and (3) "decomposability", which encourages the model to learn whole-part relations by decomposing each anatomical structure into its constituent parts. Unifying these three branches together in a coarse to fine learning approach, the localizability branch enables the model to preserve harmony in embeddings of semantically similar anatomical structures in a hierarchy of scales. Simultaneously, composability and decomposability branches empower the model to not only convey hierarchical relationships but also preserve diversity of semantically similar anatomical structures across patients through encoding finer-grained anatomical information of their constituent parts. We call our system (*i.e.*, pretrained model) **Adam–v2** because it represents a significant advancement from our previous version—Adam (**a**utodidactic **d**ense **a**natomical **m**odels) [69]—that learns *autodidactically* and yields dense anatomical embedding, nicknamed **Eve–v2** (**e**mbedding **ve**ctors) for semantic richness. We further coin our project site **Eden** (**e**nvironment for **d**ense **e**mbeddings and **n**etworks), where all code, pretrained Adam–v2, and Eve–v2 are placed.

We extensively evaluate Adam–v2 in **(1)** *Zero-shot setting* (§4.1): Adam–v2 yields more semantically meaningful embeddings (Eve–v2) compared with existing SSL methods with a set of unique properties essential for anatomy understanding (Figs. 3 to 6); **(2)** *Few-shot transfer setting* (§4.2): Adam–v2 outperforms 2 *large-scale* medical models, RadImageNet and LVM-Med, as well as a representative set of 7 SSL methods by a remarkable margin in anatomical structure and disease segmentation tasks (Tab. 1); and **(3)** *Full fine-tuning setting* (§4.3): Adam–v2 provides more generalizable representations compared to fully-supervised and SSL baselines across a myriad of tasks (Fig. 7 & Tab. 2). Our main contributions are as follows:

- A new self-supervised learning strategy, called Adam–v2, that encodes inherent hierarchical relationships within medical images, yielding discriminative representations blended with semantics of part-whole relations.
- A comprehensive set of experiments proves higher generalizability and robustness of Adam–v2, particularly highlighting Adam–v2's proficiency in few-shot transfer and achieving a new record in the ChestX-ray14 benchmark.
- A set of quantitative and qualitative feature analyses that opens up novel perspectives for assessing anatomy understanding from various viewpoints.

2. Method

Our framework, depicted in Fig. 2, aims to underpin the development of powerful self-supervised models foundational to medical imaging by constructing a hierarchy of embeddings learned from anatomy. Our framework comprises three key branches: (1) localizability, aiming to acquire discriminative representations for distinguishing different anatomical structures; (2) composability, aiming to learn each anatomical structure in a parts-to-whole manner; and (3) decomposability, aiming to comprehend each anatomical structure in a whole-to-parts manner. Seamlessly integrating these learning objectives into a unified framework captures inherent part-whole hierarchies within medical images, yielding a powerful model (Adam–v2) that can serve not only as the foundation for myriad target tasks via adaptation (fine-tuning), but also its embedding vectors (Eve–v2) bear rich semantics, usable standalone without adaptation (zero-shot), for other tasks like landmark detection. The following details our framework.

2.1. Learning Localizability

The localizability branch seeks to learn a semantically-structured embedding space where similar anatomical structures are clustered together and are distinguished from dissimilar anatomical structures. As illustrated in Fig. 2, the localizability branch includes the student g_{θ_S} and teacher g_{θ_T} encoders, and two projectors $h_{\theta_{LS}}$ and $h_{\theta_{LT}}$, referred to as localizability heads. The parameters of student g_{θ_S} and localizability head $h_{\theta_{LS}}$ are learned with stochastic gradient descent while the parameters of the teacher g_{θ_T} and head $h_{\theta_{LT}}$ are updated using an exponential moving average (EMA) on the weights of g_{θ_S} and $h_{\theta_{LS}}$, respectively. Given an anchor patch w randomly sampled from the input image I, we extract a set C of multi-scale crops from w. In particular, these crops exhibit diverse dimensions while sharing the same or slightly shifted center as w, contributing to a comprehensive understanding of the same anatomical structure at various resolutions. We then apply random data augmentations $\tau_{(.)}$ on w and multi-scale crops in C. The augmented view of w is passed to the teacher, while the augmented views of the crops in C are passed to the student network, generating the features $y_t = g_{\theta_T}(\tau(w))$ and $Y_s = \{g_{\theta_S}(\tau(c)) \mid c \in C\}$, respectively. The localizability heads project the features to the output embeddings $z_t = h_{\theta_{LT}}(y_t)$ and $Z_s = \{h_{\theta_{LS}}(y_s) \mid y_s \in Y_s\}$, which are normalized with a softmax function:

$$P_t(z_t)^{(i)} = \frac{\exp(z_t^{(i)}/\tau_t)}{\sum_{k=1}^{K} \exp(z_t^{(k)}/\tau_t)} \quad (1)$$

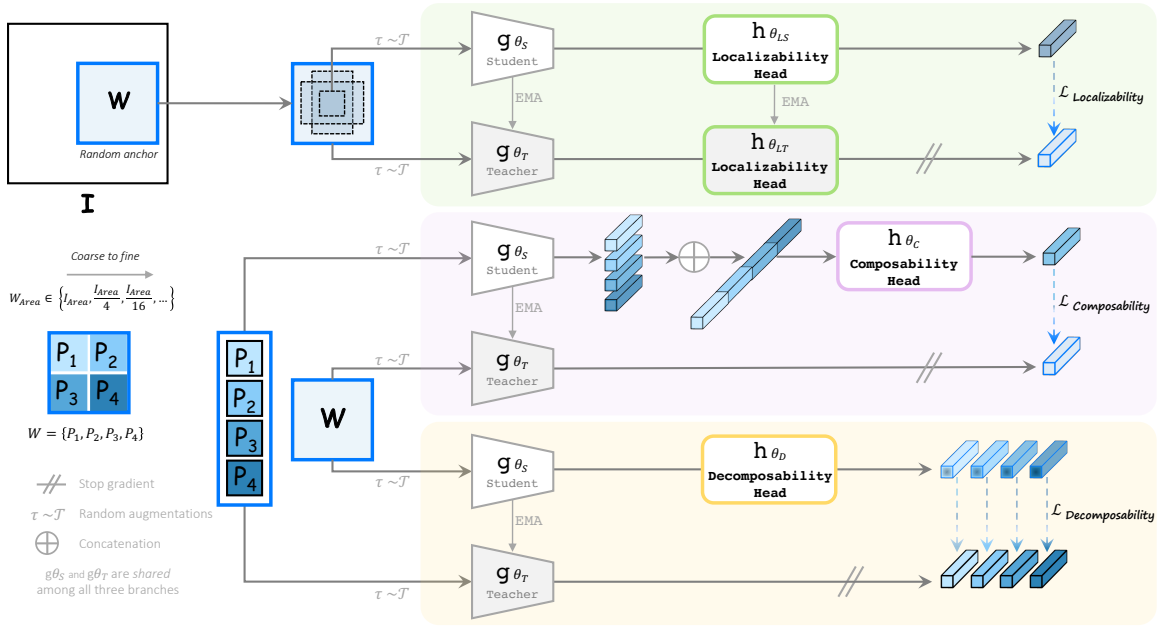

Figure 2. Adam–v2 learns hierarchical representations in a coarse-to-fine-manner via three branches: localizability, composability, and decomposability. Given an anchor whole w randomly sampled from image I, the localizability branch augment and process w and its multi-scale views, and enforce consistency between their embeddings, yielding distinct features for different anatomical structures. The composability branch decomposes w into a set of parts, and enforces consistency between the embedding of w and the aggregated embeddings of its parts, encoding *part-whole* relations. The decomposability branch decomposes the embedding of w to acquire the embeddings of its constituent parts, and enforce consistency between the embeddings of parts and their decomposed counterparts, capturing *whole-part* relations.

where $\tau_t > 0$ is a temperature parameter controlling the sharpness of the output distribution, and K is the output dimension of the localizability heads. A softmax function P_s with temperature τ_s is similarly employed to normalize the features in Z_s. The localizability branch's objective is to maximize the consistency between the embeddings of the input anchor and its augmented views. To do so, we employ cross-entropy loss [12]:

$$\mathcal{L}_{Localizability} = -\frac{1}{|Z_s|} \sum_{z_s \in Z_s} P_t(z_t) \log P_s(z_s) \qquad (2)$$

It is noteworthy that our framework offers flexibility in utilizing various localizability loss functions. While we opt for a self-distillation loss due to its simplicity and efficiency [12, 28, 67], alternative sophisticated objectives, such as contrastive loss [16, 37], can also be employed.

2.2. Learning Composability

The composability branch seeks to learn the *part-whole* anatomical hierarchies in a bottom-up manner by assembling larger anatomical structures from their smaller constituent subparts. As illustrated in Fig. 2, the composability branch consists of the student g_{θ_S} and teacher g_{θ_T} encoders, which are shared with the localizability branch, and a composability head h_{θ_C}. Given an anchor whole w randomly sampled from the input image I, we decompose it into a set of n non-overlapping parts $\mathrm{P} = \{p_i\}_{i=1}^n$. The parts are augmented and processed by the student network, generating parts' embeddings $Y_{ps} = \{y_i = g_{\theta_S}(\tau(p_i))\}_{i=1}^n$. The parts' embeddings are then concatenated and passed to the composability head h_{θ_C} to produce the aggregated embeddings of parts $z_{ps} = h_{\theta_C}(\oplus(\{y_i\}_{i=1}^n))$. Moreover, the whole anatomical structure w is augmented and passed to the teacher network to generate the whole's embedding $z_{wt} = g_{\theta_T}(\tau(w))$. The composability branch is trained to maximize the agreement between the whole's embedding and the the aggregated embeddings of its parts:

$$\mathcal{L}_{Composability} = \ell_s(z_{wt}, z_{ps}) \qquad (3)$$

where $\ell_s(z_{wt}, z_{ps})$ presents a function that measures similarity between z_{wt} and z_{ps}, such as MSE [28], cross-entropy [12], or cosine similarity [18].

2.3. Learning Decomposability

The decomposability branch seeks to learn the *whole-part* anatomical hierarchies in a top-down manner by decomposing larger anatomical structures into their smaller constituent subparts. As shown in Fig. 2, the decomposability branch comprises the student g_{θ_S} and teacher g_{θ_T} encoders,

which are shared with the localizability and composability branches, and a decomposability head h_{θ_D}. Given an anchor whole w, we decompose it into a set of n non-overlapping parts $\mathrm{P} = \{p_i\}_{i=1}^{n}$. The anchor whole w is augmented and fed into the student network, producing the whole's embedding $z_{ws} = g_{\theta_S}(\tau(w))$. The whole's embedding is then passed to the decomposability head h_{θ_D}, which decomposes it into a set of individual embeddings corresponding to the constituent parts of the whole $Z_{ps} = h_{\theta_D}(z_{ws})$. Additionally, the parts $\mathrm{P} = \{p_i\}_{i=1}^{n}$ are augmented and processed by the teacher network, generating parts' embeddings $Z_{pt} = \{g_{\theta_T}(\tau(p_i))\}_{i=1}^{n}$. The decomposability branch is trained to maximize the agreement between the embeddings of the individual parts and their decomposed counterparts:

$$\mathcal{L}_{Decomposability} = \frac{1}{|P|}\sum_{i=1}^{|P|}\ell_s(z_{p_i}, z_{p_i'}) \tag{4}$$

where $z_{p_i} \in Z_{pt}$ and $z_{p_i'} \in Z_{ps}$, and $\ell_s(z_{p_i}, z_{p_i'})$ presents a function that measures similarity between z_{p_i} and $z_{p_i'}$, such as MSE, cross-entropy, or cosine similarity.

2.4. Training Pipeline

To guide the model in learning hierarchical representations, we consider a hierarchy of diverse anatomical structures at various scales. Specifically, the highest level of the hierarchy represents entire images (of spatial resolution $(H \times W)$) with complete anatomy, while each subsequent level $m \in \{1, 2...\}$ represents anatomical structures w at a scale of $(\frac{H}{2^m} \times \frac{W}{2^m})$, randomly sampled from the images. In a coarse to fine manner, the anatomical structures w at each level are fed as the input to the localizability, composability, and decomposability branches, and are learned through the following combined loss function:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{Localizability} + \lambda_2 * \mathcal{L}_{Composability} + \lambda_3 * \mathcal{L}_{Decomposability} \tag{5}$$

where λ_1, λ_2, λ_3 are coefficients denoting the weight of each loss term. Through our unified training scheme, Adam–v2 learns a rich embedding space that preserves harmony among similar anatomical structures and encoding their hierarchical relations.

3. Implementation Details

Pretraining protocol. We use *unlabeled* chest radiographs and color fundus photographs for pretraining Adam–v2 on two imaging modalities. Our SSL framework is architecture-neutral and compatible with any ConvNet and vision transformer backbones. As an illustration, we pretrain Adam–v2 with ResNet-50 [36], ViT-S [22], and ConvNeXt-B [56] backbones. We follow [12] in optimization settings (e.g. optimizer, learning rate schedule, τ_t, τ_s, etc), updating teacher weights, and architecture of $h_{\theta_{LS}}$ and $h_{\theta_{LT}}$ heads. h_{θ_C} and h_{θ_D} are two-layer MLP heads. We use MSE as $\ell_s(.)$ in Eqs. (3) and (4). λ_1, λ_2, λ_3 are set to 1, n to 4, and m up to 4. In localizability branch, following [11, 12], we extract one 224^2 global view and eight 96^2 multi-scale crops from w. For other branches, we use input resolution 224^2. Augmentation $\tau(.)$ includes color jittering, Gaussian blur, and rotation. To prove the scalability of our framework, we train a large-scale model using ConvNeXt-B backbone and a large corpus of 926,028 images collected from 13 different public chest X-ray datasets.

Evaluations. We evaluate our framework in zero-shot, few-shot, and full transfer settings. We consider 10 downstream tasks on 9 publicly available datasets for fine-tuning settings, including JSRT [73], VinDR-Rib [61], ChestX-Det [53], SIIM-ACR [88], VinDr-CXR [62], NIH Shenzhen [46], ChestX-ray14 [76], DRIVE [9], and Drishti-GS [66]. These tasks rigorously evaluate the generalizability of our Adam–v2 across a range of applications, diseases, anatomical structures, and modalities.

Baselines. We compare Adam–v2 with a representative set of seven SOTA publicly-available SSL baselines, encompassing ConvNet- and transformer-based methods. These baselines represent diverse objectives at instance-, patch-, and pixel-level, among which TransVW [33], PCRL [94], DiRA [34], and Medical-MAE [79] represent SOTA methods tailored for medical tasks. All SSL baselines are pretrained on the same datasets as our Adam–v2 by following their official settings. Moreover, we compare Adam–v2 with the publicly available and official models of two recent *large-scale* medical models: RadImageNet [57] and LVM-Med [60], pretrained on 1.3 million medical images in fully-supervised and self-supervised manners, respectively.

Fine-tuning protocol. Following the standard transfer learning protocol [43], Adam–v2's pretrained teacher model has been fine-tuned for (1) classification tasks by appending a task-specific head, and (2) segmentation tasks by utilizing a U-Net network [65], where the encoder is initialized with the pretrained weights. We run each method for each task at least five times. We provide statistical analysis using an independent two-sample t-test.

4. Results and Analysis

4.1. Adam–v2 demonstrates zero-shot anatomy understanding, offering semantics-rich embeddings over existing SSL methods

This section showcases the anatomy understanding capabilities of our framework by delving into the unique *learned* and *emergent* properties of our Adam–v2's embeddings (Eve–v2) in various zero-shot settings.

(1) *Localazability:* We investigate Adam–v2's capability in discriminating different anatomical structures to determine

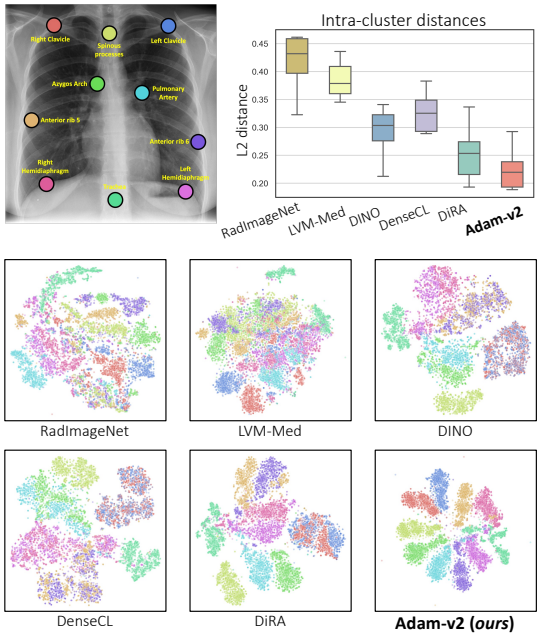

Figure 3. Adam–v2 learns localizability of anatomical structures, providing discriminative features for different landmarks. Same-colored points are instances of the same landmark across images.

if the learned embeddings (Eve–v2) preserve the locality of anatomical structures. To do so, we create a dataset of 1,000 images (from the ChestX-ray14 dataset) with 10 distinct anatomical landmarks manually annotated by human experts in each image (see Fig. 3). We extract patches of size 224^2 around each landmark's location across images and extract latent features of each landmark instance using each pretrained model under study (with *no* fine-tuning). We then visualize the embeddings with t-SNE [72] plot. We compare Adam–v2 with the RadImageNet, LVM-Med, and a representative set of SSL methods. As seen in Fig. 3, the baselines fall short in generating distinct features for different landmarks, leading to ambiguous embedding spaces with mixed clusters. By contrast, our Adam–v2 effectively discriminates between various anatomical landmarks, resulting in well-separated clusters within its learned embedding space. We complement our qualitative results (t-SNE plots) with quantitative results (box plots) by calculating intra-cluster distance for each landmark class and visualizing the distances distributions with boxplots in Fig. 3. As seen, our Adam–v2 exhibits lower median distances, indicating more cohesive clusters, compared to the baselines. To showcase Adam–v2's capacity in balancing anatomical diversity and harmony and conveying hierarchical relationships, we randomly select four distinct anatomical landmarks, extract three patches of different resolutions (labeled as levels 1, 2, and 3) around each landmark across the images, and compute their embeddings with Adam–v2's

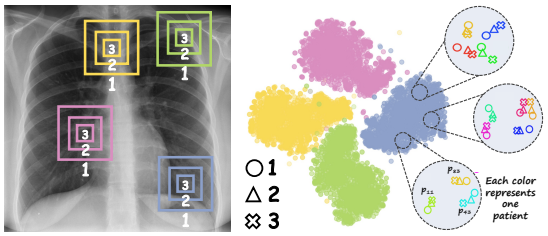

Figure 4. Adam–v2 balances diversity and harmony in embeddings of similar anatomical structures across patients and scales.

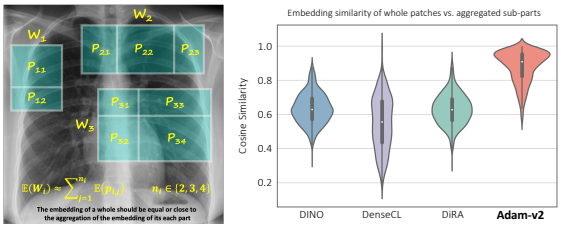

Figure 5. Adam–v2's embeddings (Eve–v2) encode part-whole relations of anatomical structures.

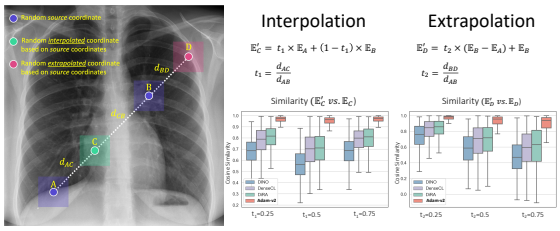

Figure 6. Adam–v2 exhibits two emergent properties: Interpolation & Extrapolation. For interpolation/extrapolation, similarity has been computed between the interpolated/extrapolated embeddings (E'_C/E'_D) and their corresponding ground truth (E_C/E_D).

pretrained model. As seen in Fig. 4, the embeddings of anatomical structures at levels 1, 2, and 3 for each landmark are closely aligned, highlighting Adam–v2's capability to preserve harmony in embeddings of semantically similar anatomical structures across resolutions and patients. Also, within each landmark, the embeddings of patches with levels 1, 2, and 3 for the same patient (color-coded in Fig. 4) are close, while those of different patients are well separated, representing Adam–v2's capability to preserve diversity of anatomical structures across patients.

(2) *Composability & Decomposability*: We explore Adam–v2's ability to capture part-whole hierarchies, as imposed by the composability and decomposability branches, in its learned embeddings (Eve–v2). To do so, we extract random patches of varying sizes, called *whole*, from ChestX-ray14 test images. Each *whole* is decomposed into 2, 3, or 4 non-overlapping parts with different sizes. We resize each *whole* and its parts to 224^2, extract features using pretrained mod-

Method	Anatomical Structure Segmentation								Disease Segmentation			
	JSRT-Clavicle (Dice%)				JSRT-Heart (Dice%)				SIIM-ACR (Dice%)		ChestX-Det (IoU%)	
	3-shot	6-shot	12-shot	24-shot	3-shot	6-shot	12-shot	24-shot	5%	10%	5%	10%
RadImageNet [57]	55.52	71.26	82.57	83.29	73.12	75.42	89.22	91.00	54.56	61.48	64.22	67.10
LVM-Med [60]	56.87	72.99	83.48	84.10	79.45	86.94	89.98	90.78	54.13	62.31	65.11	67.14
DINO [12]	24.06	29.59	38.54	45.01	45.45	60.79	70.85	80.78	47.85	52.08	46.84	52.64
DenseCL [77]	36.43	51.31	63.03	69.13	64.88	74.43	75.79	80.06	48.07	52.32	60.18	65.76
DiRA [34]	31.42	38.59	66.81	73.06	63.76	64.47	76.10	81.42	42.44	48.27	61.63	64.86
Adam–v2 (Ours)	**73.59**	**79.57**	**84.00**	**85.96**	**86.88**	**89.87**	**90.47**	**91.39**	**55.61**	**68.11**	**65.92**	**68.17**
Δ_1	+16.7	+6.58	+0.52	+1.86	+7.43	+2.93	+0.49	+0.39	+1.05	+6.50	+0.81	+1.03
Δ_2	+37.1	+28.2	+17.2	+12.9	+22.0	+15.4	+14.3	+9.97	+7.54	+15.7	+4.29	+2.41

Table 1. Adam–v2 excels in few-shot transfer, outperforming large-scale medical models (RadImageNet and LVM-Med) and SSL baselines across segmentation tasks. Δ_1 and Δ_2 show Adam-v2's performance boosts over second-best large-scale and SSL baselines, respectively.

els, and calculate the cosine similarity between the embedding of each *whole* and the aggregate of its parts. As seen in Fig. 5, the box plot elements indicate that the median similarity for our Adam–v2 is significantly higher than that of other SSL baselines. Additionally, the distribution of our Adam–v2's similarity values is highly concentrated around the 1.5x interquartile, situated at the top of the box plot. This concentration suggests that, in most cases, the similarity value between the embedding of entire *wholes* and their aggregated parts is closer to 1 in our Adam–v2 model.

(3) *Interpolatation* & (4) *extrapolation*: We investigate the Adam–v2's capability to interpolate/extrapolate embeddings for a randomly chosen anatomical structure by leveraging the embeddings of two other randomly selected anatomical structures. For interpolation, we select two random source coordinates (labeled as A and B in Fig. 6) and use the established interpolation formula (refer to Fig. 6) to interpolate a random point C. We extract 224^2 patches around points A, B, and C and pass them through each pretrained model under study to extract their respective embeddings E_A, E_B, and E_C, where E_C serves as the ground truth for evaluating the interpolated embeddings for C. Subsequently, we apply the interpolation formula to generate embeddings for C based on E_A and E_B, resulting in interpolated embeddings E'_C. Finally, we compute the cosine similarity between the interpolated embeddings E'_C and the ground truth E_C. This process was repeated for 1,000 images selected from the test images of Chest X-ray 14, employing three different values of t_1 (i.e., 0.25, 0.5, and 0.75). We use boxplots to illustrate the similarity distributions in each setting. We examine extrapolation of embeddings for a randomly selected point D in a similar manner using the extrapolation formula. The boxplots in Fig. 6 reveal the consistent superiority of our Adam–v2 in delivering higher similarity between interpolated/extrapolated embeddings and the ground truth (with a median close to 1) compared to other baselines. This outstanding performance is indicative of the Adam–v2's capability in establishing relations between anatomical structures. It's noteworthy that our Adam–v2 model was *not* explicitly trained for these properties, and their emergence underscores the Adam–v2's capabilities in understanding anatomy.

4.2. Adam–v2 excels in few-shot transfer, outperforming SOTA fully/self-supervised methods in segmentation tasks

This section highlights the effectiveness of Adam–v2 as an effective foundation for fine-tuning deep models in segmentation tasks with limited labeled data. We compare Adam–v2 with 3 SSL methods, as well as RadImageNet and LVM-Med models, which serve as performance upper bounds. We conduct experiments on heart and clavicle segmentation tasks, fine-tuning the pretrained models using a few shots of labeled data (3, 6, 12, and 24) randomly sampled from JSRT dataset. Moreover, we conduct experiments on various thoracic disease segmentation tasks, fine-tuning the pretrained models on two randomly selected label fractions (5% and 10%) of the SIIM-ACR and ChestX-Det datasets. As seen in Tab. 1, our Adam–v2 outperforms both RadImageNet and LVM-Med across all label fractions in all tasks. For instance, in the 3-shot transfer for clavicle and heart segmentation tasks, Adam–v2 surpasses LVM-Med by at least 16% and 7%, respectively. Moreover, Adam–v2 provides outstandingly better few-shot transfer performance compared with SSL methods across all tasks. For instance, in the pneumothorax segmentation task within the SIIM-ACR dataset, our Adam–v2 surpasses the runner-up baseline by 7.54% and 15.7% in the 5% and 10% labeled data subsets, respectively. Similarly, across the 5% and 10% fractions of the ChestX-Det dataset, our Adam–v2 demonstrates notably higher averages of 4.29% and 2.41% in the thoracic diseases segmentation task. Our attribution of Adam–v2's superior representations for few-shot segmentation tasks is grounded in the significance of anatomy learning through our SSL approach and its profound impact on representation learning, which is neglected in existing methods.

4.3. Adam–v2 stands out in full transfer, unleashing generalizable representations for a variety of tasks

This section demonstrates the generalizability of Adam–v2's representations via transfer learning to a broad range of

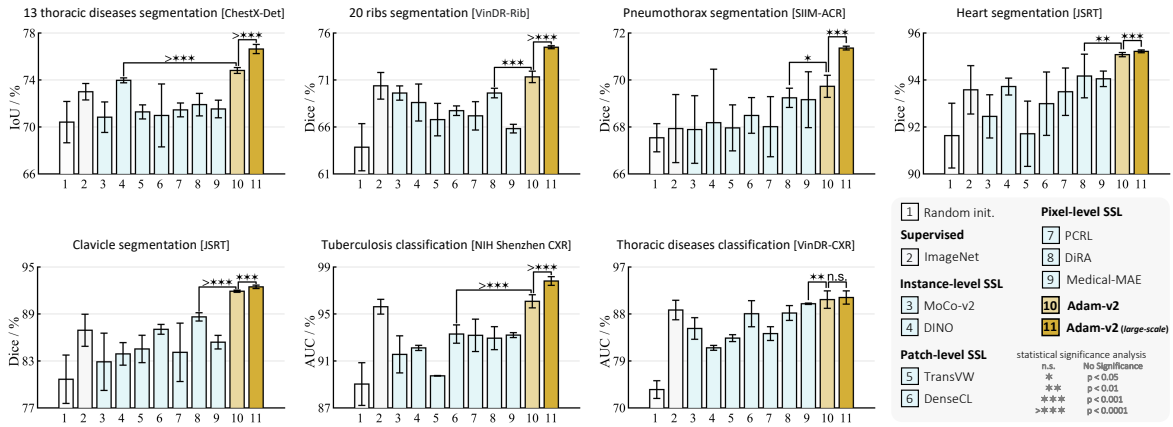

Figure 7. Adam–v2 provides generalizable and robust representations, outperforming SOTA self-supervised methods across diverse downstream tasks. Statistical significance analysis ($p < 0.05$) was conducted between Adam–v2 and the top SSL baseline in each task.

Method	# Pretraining Data	AUC†
RadImageNet [57]	1.3M	80.7
LVM-Med [60]	1.3M	82.0
Medical MAE [79]	0.5M	83.0‡
Adam–v2 (*Large-scale*)	∼1M	**83.4**

† We report mean AUC over 14 diseases on the official test split of ChestX-ray14 dataset.
‡ We adopted this performance reported by the original authors [79]; All the rest performance is ours.

Table 2. Adam–v2 outperforms previous SOTA methods (officially released large-scale medical vision models) on the public ChestX-ray14 benchmark, yielding a new record mAUC of 83.4%.

downstream tasks in a full fine-tuning setting. We compare Adam–v2 with 7 SOTA ConvNet- and vision transformer-based SSL methods designed for both computer vision and medical applications. We include training downstream models from random initialization (the lower-bound baseline) and fully-supervised ImageNet model. As seen in Fig. 7, our Adam–v2 consistently achieves superior performance compared with the fully-supervised ImageNet model, as well as significant performance boosts ($p < 0.05$) compared with all SSL counterparts across all tasks.

Comparison in Public ChestX-ray14 Benchmark. To scrutinize the scalability of our framework, we pretrained Adam–v2 with the ConvNeXt-B backbone on nearly 1M chest X-ray images and compared it against officially released large-scale medical vision models in the ChestX-ray14 benchmark. As seen in Tab. 2, Adam–v2 hits a new record of 83.4 in the ChestX-ray14 benchmark. This suggests that a meticulously crafted learning strategy that comprehends human anatomy can fully harness large-scale data, thereby paving the way for developing powerful self-supervised models foundational to medical imaging.

4.4. Ablation Experiments

Generalizability of our framework. Our framework can seamlessly extend to other imaging modalities. To demonstrate this, we consider fundus images and pretrain Adam–v2 using the EyePACS dataset and then fine-tune it for two downstream tasks, considering both low-data regimes and full fine-tuning settings. As seen in Tab. 3, Adam–v2 exhibits superior performance ($p < 0.05$) across tasks in both settings compared with SSL baselines that leverage the same pretraining data as our Adam–v2. Moreover, Adam–v2 outperforms ($p < 0.05$) RadImageNet and LVM-Med models in low-data regimes and achieves superior or equivalent performance in full fine-tuning scenarios.

Effect of learning objectives. We assess the impact of each learning branch in Adam–v2 by starting from localizability and incrementally adding composability and decomposability learning. We fine-tune the models for two downstream tasks. As seen in the top-row of Fig. 8, augmenting localizability with composability learning consistently improves performance across tasks. Moreover, the inclusion of decomposability further enhances the performance, resulting in significant performance boosts ($p < 0.05$) in both tasks compared to standalone localizability learning.

Effect of coarse-to-fine learning. We investigate the impact of hierarchical learning of anatomical structures at various scales (i.e. m) by initially training Adam–v2 with the entire anatomy ($m = 0$) and then progressively delving deeper into the higher levels of anatomy hierarchy (up to level 3), representing finer anatomical structures. As seen in bottom-row of Fig. 8, gradual increment of data granularity from $m = 0$ to $m = 2$ consistently improves the downstream performance. This highlights that our coarse-to-fine learning strategy incrementally deepens the model's anatomical knowledge, resulting in more generic representations for

Method	DRIVE (Dice%)		Drishti-GS (Dice%)	
	10%	100%	10%	100%
Random	74.03 (0.87)	78.27 (0.40)	70.17 (10.91)	94.53 (1.72)
RadImageNet	76.53 (0.49)	78.55 (0.17)	90.37 (1.48)	96.33 (0.15)
LVM-Med	77.19 (0.75)	79.46 (0.14)	91.60 (2.19)	**97.02 (0.15)**
DINO	75.89 (0.63)	78.36 (0.28)	85.90 (3.27)	96.44 (0.33)
DenseCL	75.76 (0.90)	78.36 (0.47)	86.04 (3.27)	96.60 (0.01)
DiRA	75.92 (0.90)	78.52 (0.38)	91.19 (1.86)	96.76 (0.16)
Adam–v2 (Ours)	**78.04 (0.14)***	**79.91 (0.18)***	**94.04 (0.56)***	**97.02 (0.19)**

Table 3. Adam–v2 outperforms SSL methods in fundus downstream tasks. $\star$ shows statistically significant ($p < 0.05$) boosts.

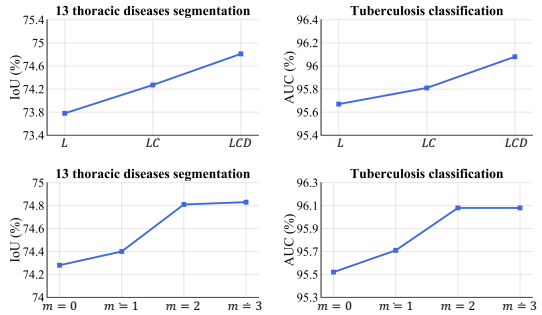

Figure 8. Ablation on the impact of (a) different branches of Adam–v2 (top-row) and (b) coarse-to-fine learning (bottom-row).

myriad tasks. Additionally, no significant change in performance is observed at $m = 3$, suggesting that pretraining up to level 2 yields sufficiently robust representations.

5. Related Work

Self-supervised learning. A large body of SSL methods seek to learn global features via instance discrimination pretext tasks. These methods align the features of augmented views from the same image by employing diverse learning objectives, including *contrastive learning* [11, 16, 17, 19–21, 29, 37, 52, 87, 90], *self-distillation* [10, 12, 12, 18, 27, 28, 67], and *feature decorrelation* [5, 7, 23, 89, 91]. Alternatively, dense SSL methods seek to learn local features by encoding visual patterns embedded at smaller image regions. *Dense contrastive learning* methods [78, 86, 93] enforce consistency between pixels at the same spatial location [6, 63, 82], similar pixels/patches in a feature map [6, 77], or similar image regions [80, 81, 85, 92]. On the other hand, *masked image modeling* methods [4, 14, 24, 38, 48, 50, 54, 58, 71, 74, 75, 83, 84] mask random portions of the images and reconstruct the missing parts at pixel-level. Motivated by the success in computer vision, a broad variety of instance discrimination [2, 3, 49] and image reconstruction methods [15, 79], along with their integration [34, 44, 70, 94], have been explored for medical imaging. Given such advancements, the evolution of SSL has empowered it to serve as the cornerstone for developing foundation models with broad applicability [8]. However, existing SSL methods overlook anatomy hierarchies in their learning objectives, thereby lacking anatomy understanding capabilities. By contrast, Adam–v2 exploits the hierarchical nature of anatomy to learn semantics-rich features, leading to more pronounced models tailored for medical tasks.

Learning from anatomy. Consistent anatomy in medical imaging provides strong yet free supervision signals for deep models to learn common anatomical representations via self-supervision [95]. Existing works revolve around recovering anatomical patterns from transformed images [95, 96], learning semantics of recurrent anatomical patterns across patients [32, 33] with subsequent enhancements via adversarial learning [30, 31, 34, 35], exploiting spatial relationships in anatomy [64], utilizing global and local anatomical consistency [97], and incorporating anatomical cues to improve contrastive learning [13, 25, 45, 47]. These existing works neglect hierarchical anatomy relations. Although our earlier method Adam [69] uses anatomy hierarchies as soft supervisory signals, our Adam–v2 explicitly encodes part-whole hierarchies via its learning objectives. Compared with Adam [69], Adam–v2 showcases two significant advancements: (1) enhancing the localizability branch by eliminating negative pairs pruning, thereby improving computational efficiency for large-scale pretraining, (2) introducing two novel components: composability and decomposability, which are crucial for capturing part-whole hierarchies.

Learning part-whole hierarchies. Hierarchical representation learning is ingrained in architectures such as ConvNets [36, 56] and hierarchical vision transformers (ViT) [55]. But, the multi-scale feature hierarchy of common neural networks does not explicitly align with the part-whole hierarchy in images, leading to the advent of new architectures for encoding part-whole hierarchies [42, 51]. Notably, GLOM [39] introduced a conceptual framework that utilizes attention to learn part-whole hierarchies, and subsequent works proposed ViT-based architectures to implement it [26, 68]. By contrast, Adam–v2 goes beyond architecture design by introducing a new learning strategy that encodes the semantics of part-whole hierarchies into the embedding space through three explicit training objectives: localizability, composability, and decomposability.

6. Conclusion

We present a SSL framework Adam–v2 that enhances visual representations by creating a hierarchy of embeddings for different anatomical structures. The major novelty of our work is *explicitly* enforcing part-whole hierarchies via three learning objectives. Our experiments highlight the effectiveness of Adam–v2 in various tasks, surpassing a range of baselines. We also demonstrate the semantic richness of our learned representations, which stem from explicitly acquired or autonomously emerging unique properties.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022. 1

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468, 2021. 8

[3] Shekoofeh Azizi, Laura Culp, Jan Freyberg, and et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7:756–779, 2023. 8

[4] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24256–24265, 2023. 8

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021. 8

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features, 2022. 8

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 8

[8] Rishi Bommasani and et al. On the opportunities and risks of foundation models, 2021. 8

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

[10] Yun-Hao Cao, Peiqin Sun, and Shuchang Zhou. Three guidelines you should know for universally slimmable self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15742–15751, 2023. 8

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924. Curran Associates, Inc., 2020. 4, 8

[12] 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. 3, 4, 6, 8

[13] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, pages 12546–12558. Curran Associates, Inc., 2020. 8

[14] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22742–22751, 2023. 8

[15] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 8

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 3, 8

[17] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020. 8

[18] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 3, 8

[19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 8

[20] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021.

[21] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16670–16681, 2022. 8

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[23] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3015–3024. PMLR, 2021. 8

[24] Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10386–10395, 2023. 8

[25] Zeyu Fu, Jianbo Jiao, Robail Yasrab, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Anatomy-aware contrastive representation learning for fetal ultrasound. In *Computer Vision – ECCV 2022 Workshops*, pages 422–436, 2023. 8

[26] 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 Vi-*

sion and Pattern Recognition (CVPR), pages 13689–13698, 2022. 8

[27] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, and P. Perez. Obow: Online bag-of-visual-words generation for self-supervised learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6826–6836, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 8

[28] 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. 3, 8

[29] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9706–9715, 2022. 8

[30] Zuwei Guo, Nahid U.I. Islam, Michael B. Gotway, and Jianming Liang. Discriminative, restorative, and adversarial learning: Stepwise incremental pretraining. In *Domain Adaptation and Representation Transfer*, pages 66–76, 2022. 8

[31] Zuwei Guo, Nahid Ul Islam, Michael B. Gotway, and Jianming Liang. Stepwise incremental pretraining for integrating discriminative, restorative, and adversarial learning. *Medical Image Analysis*, page 103159, 2024. 8

[32] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 137–147, Cham, 2020. Springer International Publishing. 8

[33] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 40(10):2857–2868, 2021. 4, 8

[34] 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. 4, 6, 8

[35] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? *Medical Image Analysis*, 94:103086, 2024. 8

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 8

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 8

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 8

[39] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35(3):413–452, 2023. 1, 8

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

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

[42] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *International Conference on Learning Representations*, 2018. 8

[43] 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. 4

[44] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B. Gotway, and Jianming Liang. 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*, pages 535–551. PMLR, 2022. 8

[45] Hongyu Hu, Tiancheng Lin, Yuanfan Guo, and et al. Anatomy-aware self-supervised learning for aligned multimodal medical data. In *British Machine Vision Conference*, 2022. 8

[46] 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

[47] Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15859–15869, 2023. 8

[48] Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, and Zhangyang Wang. Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[49] Aakash Kaku, Sahana Upadhya, and Narges Razavian. Intermediate layers matter in momentum contrastive self super-

vised learning. In *Advances in Neural Information Processing Systems*, pages 24063–24074. Curran Associates, Inc., 2021. 8

[50] Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7918–7928, 2023. 8

[51] Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked capsule autoencoders. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 8

[52] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021. 8

[53] 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. 4

[54] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6252–6261, 2023. 8

[55] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 8

[56] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 8

[57] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022. 4, 6, 7

[58] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations, 2022. 8

[59] 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

[60] Duy M. H. Nguyen, Hoang Nguyen, Nghiem T. Diep, Tan N. Pham, Tri Cao, Binh T. Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, Daniel Sonntag, and Mathias Niepert. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching, 2023. 4, 6, 7

[61] 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. 4

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

[63] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *Advances in Neural Information Processing Systems*, pages 4489–4500. Curran Associates, Inc., 2020. 8

[64] Jiaxuan Pang, Fatemeh Haghighi, DongAo Ma, Nahid Ul Islam, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. POPAR: Patch order prediction and appearance recovery for self-supervised medical image analysis. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 77–87. Springer, 2022. 8

[65] 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. 4

[66] 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

[67] Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multi-mode online knowledge distillation for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11848–11857, 2023. 3, 8

[68] Shuyang Sun, Xiaoyu Yue, Song Bai, and Philip Torr. Visual parser: Representing part-whole hierarchies with transformers, 2022. 8

[69] 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. 1, 2, 8

[70] Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, 2022. 8

[71] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, 2023. 8

[72] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 5

[73] 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. 4

[74] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10375–10385, 2023. 8

[75] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2122–2131, 2023. 8

[76] 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. 4

[77] 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. 6, 8

[78] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16590–16599, 2022. 8

[79] 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. 4, 7, 8

[80] Tete Xiao, Colorado J Reed, Xiaolong Wang, Kurt Keutzer, and Trevor Darrell. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10539–10548, 2021. 8

[81] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8392–8401, 2021. 8

[82] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693, 2021. 8

[83] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 8

[84] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10365–10374, 2023. 8

[85] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Regioncl: Exploring contrastive region pairs for self-supervised representation learning. In *Computer Vision – ECCV 2022*, pages 477–494, Cham, 2022. Springer Nature Switzerland. 8

[86] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3987–3996, 2021. 8

[87] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

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

[89] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230*, 2021. 8

[90] Chaoning Zhang, Kang Zhang, Trung X. Pham, Axi Niu, Zhinan Qiao, Chang D. Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14441–14450, 2022. 8

[91] Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2022. 8

[92] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[93] Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16580–16589, 2022. 8

[94] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3499–3509, 2021. 4, 8

[95] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*

2019, pages 384–393, Cham, 2019. Springer International Publishing. 8

[96] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. 8

[97] 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. 1, 8