Learning Anatomically Consistent Embedding for Chest Radiography

Ziyu Zhou*[1,2]
zhouziyu@sjtu.edu.cn
Haozhe Luo*[2]
hluo54@asu.edu
Jiaxuan Pang[2]
jpang12@asu.edu
Xiaowei Ding†[1]
dingxiaowei@sjtu.edu.cn
Michael Gotway[3]
gotway.michael@mayo.edu
Jianming Liang†[2]
Jianming.Liang@asu.edu

[1] Shanghai Jiao Tong University, China
[2] Arizona State University, USA
[3] Mayo Clinic, USA

Abstract

Self-supervised learning (SSL) approaches have recently shown substantial success in learning visual representations from unannotated images. Compared with photographic images, medical images acquired with the same imaging protocol exhibit high consistency in anatomy. To exploit this anatomical consistency, this paper introduces a novel SSL approach, called PEAC (patch embedding of anatomical consistency), for medical image analysis. Specifically, in this paper, we propose to learn global and local consistencies via stable grid-based matching, transfer pre-trained PEAC models to diverse downstream tasks, and extensively demonstrate that (1) PEAC achieves significantly better performance than the existing state-of-the-art fully/self-supervised methods, and (2) PEAC captures the anatomical structure consistency across views of the same patient and across patients of different genders, weights, and healthy statuses, which enhances the interpretability of our method for medical image analysis. All code and pretrained models are available at GitHub.com/JLiangLab/PEAC.

1 Introduction

Self-supervised learning (SSL) [18] pretrains generic source models without using expert annotation, allowing the pretrained generic source models to be quickly fine-tuned into high-performance application-specific target models [38] and minimizing annotation cost [24]. This paradigm is particularly attractive in medical imaging because labeling medical images is tedious, laborious, and time-consuming and demands specialty-oriented expertise [25, 37]. However, most existing SSL methods were developed for photographic images, and directly

∗ Equal contribution. † Corresponding author.

adopting these SSL methods to medical images may not achieve optimal results because medical images are markedly different from photographic images. Photographic images, like those in ImageNet [9], are object-centric, where dominant objects (e.g., dogs and cats) are located at the center with backgrounds of large variation. Naturally, these SSL methods developed for photographic images mostly learn from foreground objects. By contrast, medical images acquired with the same imaging protocol have similar anatomical structures, and imaging diagnosis requires not only analyzing "foreground" objects—diseases, but also understanding "background" objects—anatomical structures. Furthermore, diseases are often small and obscured in "background" anatomical structures. Therefore, we hypothesize that SSL achieves better performance in medical imaging when learning from anatomy in health and disease. To test this hypothesis, we have chosen chest X-rays because the chest contains several critical organs prone to a number of diseases associated with significant healthcare costs [38], and chest X-rays are one of the most frequently used modalities in imaging the chest. In chest X-rays, as illustrated in Fig. 1, there are large and small anatomical structures, such as the right/left lung, heart, and spinous processes; lung diseases can be local or global. This paper seeks to answer this critical question: *How to autodidactically learn generic source models from global and local patterns in health and disease?*

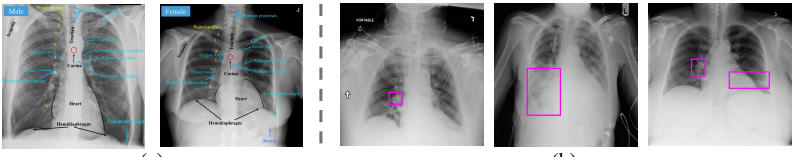
To answer this question, we have developed a novel SSL framework, called PEAC (patch embedding of anatomical consistency), to exploit global and local patterns in health and disease. As illustrated in Fig. 2-3, PEAC has an architecture of student-teacher, taking two global crops, one for the student and the other for the teacher, with overlaps from a chest X-ray to learn the global consistency between the two global crops and the local consistency between each pair of corresponding local patches within the overlapped region of the two global crops. Our extensive experiments have demonstrated that our PEAC outperforms fully-supervised pretrained models on ImageNet or ChestX-ray14 and SoTA SSL methods, and offers consistent representation of similar anatomical structures across diverse patients of different genders and weights and across different views of the same patient.

Through this work, we have made the following contributions:

- A straightforward yet effective SSL scheme that captures both global and local patterns embedded within medical images;
- A precise and stable patch-matching method that achieves anatomical embedding consistency in health and disease;
- Extensive illustrations that show the capability of PEAC in matching anatomical structures across different patients and across different views of the same patient and in segmenting anatomical structures by zero-shot;
- Thorough experiments that demonstrate the transferability of PEAC to various target tasks, outperforming SoTA full-supervised and self-supervised methods in classification and segmentation.

2 Related Work

Global features and local features. Global features describe the overall appearance of the image. Most recent methods for global feature learning are put forward to ensure that the extracted global features are consistent across different views. The methods to achieve



(a) (b)

Figure 1: (a) Chest X-rays contain various large (global) and small (local) anatomical patterns, including the right/left lung, heart, spinous processes, clavicle, mainstem bronchus, hemidiaphragm, and the osseous structures of the thorax, that can be utilized for learning consistent embedding in anatomy. (b) Diagnosing chest diseases at chest X-rays involves identifying focal and diffuse patterns, such as Mass, Infiltrate, and Atelectasis as boxed, that can be exploited for learning consistent embedding in disease.

this include contrastive learning and non-contrastive learning methods. Contrastive methods [12, 14, 15, 20, 21, 27] bring representation of different views of the same image closer and spreading representations of views from different images apart. Non-contrastive methods rely on maintaining the informational content consistent of the representations by either explicit regularization [3, 9, 36] or architecture design like Siamese architecture [6, 10, 19]. In opposition to global methods, local features describe the information that is specific to smaller regions of the image. In local features learning methods, a contrastive or consistent loss can be applied directly at the pixel level [32], the feature map level [9, 29] or the image region level [31] which forces consistency between pixels at similar locations, between groups of pixels and between large regions that overlap in different views of an image. However, at present the vast majority of methods that use local features calculate embedding consistency or contrastive learning loss based on the relative positions of the features [4, 34, 35], such as the feature vectors of semantically closest patches or spatially nearest neighbor patches. In contrast, our PEAC method calculates the consistency loss based on the absolute positions of overlapping image patches shown in Fig. 2. In this way, fine-grained anatomical structures can be more accurately characterized.

3 Our Method

The goal of our method, Patch Embedding of Anatomical Consistency (PEAC), is to learn global and local anatomical structures underneath medical images. In medical images, there are various local patterns such as spinous processes, clavicle, mainstem bronchus, hemidiaphragm, the osseous structures of the thorax, etc. The resemblances can be captured by the two global crops shown in Fig. 3 so that global embedding consistency can encourage the network to extract high-level semantic features of similar local regions. Besides, local embedding consistency based grid-like image patches can equip the model with a more stable matching strategy, for disease diagnosing which needs both single and multiple local patterns to catch the fine-grained anatomical structure. Therefore, we proposed a network that considers both global and local features of medical images at the same time.

As shown in Fig. 3, PEAC is an SSL framework comprised of four key components: (1) Student-Teacher model that aims to extract features of two crops simultaneously; (2) image augmentation and restoration module aim to restore image crops from the patch order and appearance distortion; (3) global module that aims to enforce the model to learn coarse-grained

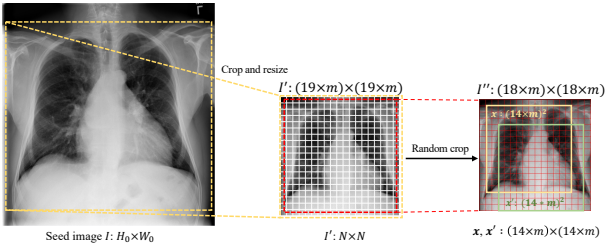


Figure 2: Grid-wise cropping for stable grid-based matching. Firstly, a seed image I is cropped from an original chest X-ray and resized to Image I' of size $(n\times m)\times(n\times m)$, so that I' can be conveniently partitioned to $n\times n$ patches with each patch of size $m\times m$. By default, $n=19$ and $m=32$ in PEAC. Secondly, Image I'' with size $((n-1)\times m)\times((n-1)\times m)$ is randomly cropped from I' to ensure a large diversity of local patches during training. Thirdly, Crops x and x' of size $(k\times m)\times(k\times m)$ are randomly extracted from Image I'' in alignment with the grid of Image I'' to ensure the exact correspondence of local matches in the overlapped region between Crops x and x' (referred to stable grid-based matching and detailed in Sec. 4.3). By default, $k=14$ in PEAC. See the Supplementary Materials for various PEAC configurations.

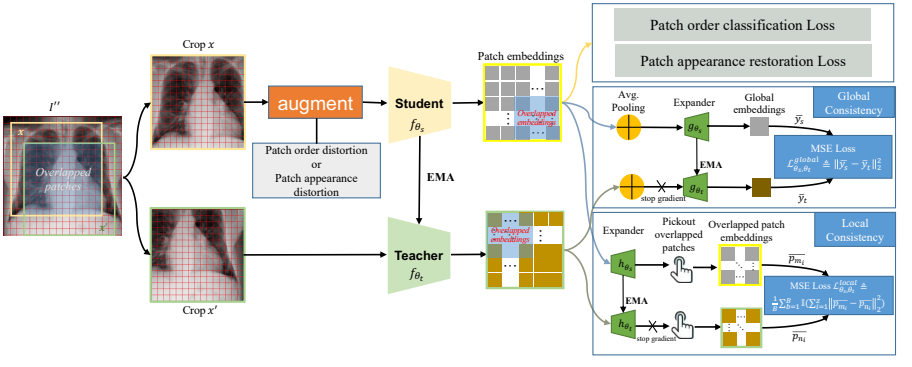


Figure 3: PEAC has an architecture of student-teacher, taking two global crops: x for the student and x' for the teacher, with overlaps from a chest X-ray to learn the global consistency (Eq. 1) between the two global crops (x and x') and the local consistency (Eq. 2) between each pair of corresponding local patches within the overlapped region of the two global crops (x and x'). The student, built on POPAR [22], learns high-level relationships among anatomical structures by patch order classification and fine-grained image features by patch appearance restoration, as detailed in the Supplementary Materials. Integrating the teacher with the student aims to learn consistent contextualized embedding for coarse-grained global anatomical structures and fine-grained local anatomical structures across different views of the same patients, leading to anatomically consistent embedding across patients.

global features of two crops; (4) local module that aims to enforce the model to learn fine-grained local features from overlapped patches. By integrating the above modules, the model learns the coarse-grained, fine-grained and contextualized high-level anatomical structure features. In the following, we will introduce our methods from image pre-processing, each components and the joint training loss. Our model is based on POPAR because we need to model the overall structural information and local detailed and robust information of medical

images.

3.1 Global Embedding Consistency

Before inputting to the model, the seed images are pre-processed in grid-wise cropping to get two crops x, $x' \in R^{C\times H\times W}$, C is the number of channels, (H,W) are the crops' spatial dimensions, shown in Fig. 2. Then the two crops are input to the *Student* and *Teacher* encoders $f_{\theta_s}, f_{\theta_t}$ to get the local features s, t respectively. Then in the global branch the average pooling operators $\oplus : R^{D\times H\times W} \rightarrow R^D$ are performed on the local features. We denote the pooled representations as $y_{s\oplus}$ and $y_{t\oplus} \in R^D$. At last the expanders g_{θ_s}, g_{θ_t} are 3-layer MLP which map $y_{s\oplus}$, $y_{t\oplus}$ to get the embedding vectors $y_s, y_t \in R^H$. We put the l_2-normalize to $\overline{y_s} = y_s/\|y_s\|_2$ and $\overline{y_t} = y_t/\|y_t\|_2$. . At last, we define global patch embedding consistency loss as the following mean square error between the normalized output,

$$\mathcal{L}^{global}_{\theta_s,\theta_t} \triangleq \|\overline{y_s} - \overline{y_t}\|_2^2 = 2 - 2\cdot\frac{\langle y_s, y_t\rangle}{\|y_s\|_2\cdot\|y_t\|_2} \tag{1}$$

We symmetrize the loss from Eq. 1 by separately feeding x to *Teacher* encoder and x' to *Student* encoder to compute $\widetilde{\mathcal{L}}^{global}_{\theta_s,\theta_t}$. Accordingly, we get the global loss as $\mathcal{L}^{G}_{\theta_s,\theta_t} = \mathcal{L}^{global}_{\theta_s,\theta_t} + \widetilde{\mathcal{L}}^{global}_{\theta_s,\theta_t}$.

3.2 Local Embedding Consistency

As the encoders are Vision Transformer network, the crop is divided into a sequence of N non-overlapping image patches $P = (p_1, p_2, ..., p_N)$ where $N = \frac{H\times W}{m^2}$ and m is the patch resolution. The encoder of the Student-Teacher model extracts local features $s,t \in R^{D\times N}$ from the two crops x,x'. We denote s_k and $t_k \in R^D$ the feature vectors at position $k \in [1,...,N]$ in their corresponding feature maps. Since the image patches are randomly sampled from an image grid with an overlap rate of 50%-100%, we define the overlapping image patches O_m, O_n for x and x' respectively, and $m \in [m_1,...,m_z]$, $n \in [n_1,...,n_z]$ are the patch indexes of the overlapping region, z is the number of overlapping patches. O_{m_i} and O_{n_i} are in correspondece where $1 \le i \le z$ and we call this process grid matching. Correspondingly, O_m and O_n are transformed into embedding vectors o_m and o_n through the feature extractors. Then in the local module there are 3-layer MLP expanders h_{θ_s}, h_{θ_t} adding to o_m, o_n to get the final local patch embedding vectors p_m, p_n. Similarly, we also put l_2-normalize to $\overline{p_m} = p_m/\|p_m\|_2$, $\overline{p_n} = p_n/\|p_n\|_2$. We only randomly add patch order distortion and patch appearance distortion in the student branch. When the patch order is distorted, the patch embedding vector will represent the distorted global feature for attention mechanism. And local embeddings of distorted and non-distorted patch orders in the student and teacher branches can't be consistent. So we won't compute local loss if the crop gains patch order distortion (indicator $\mathbb{I} = 0$) while it has no impact on the patch appearance distortion ($\mathbb{I} = 1$). To align the output of the student and teacher networks regarding local features, we define the following local patch embedding consistency loss function in Eq. 2

$$\mathcal{L}^{local}_{\theta_s,\theta_t} \triangleq \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\cdot(\sum_{i=1}^{z}\|\overline{p_{m_i}} - \overline{p_{n_i}}\|_2^2) \tag{2}$$

$\overline{p_{m_i}}$ and $\overline{p_{n_i}}$ are the embedding vectors of the i-th overlapping image patches and B is the batch size. Similar to the global loss in previous section, when x is fed into *Teacher* encoder

and x' is fed into *Student* encoder, we compute corresponding loss $\widetilde{\mathcal{L}}^{local}_{\theta_s,\theta_t}$. So the local loss $\mathcal{L}^{L}_{\theta_s,\theta_t} = \mathcal{L}^{local}_{\theta_s,\theta_t} + \widetilde{\mathcal{L}}^{local}_{\theta_s,\theta_t}$.

We calculate $\mathcal{L}^{oc}_{\theta_s} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{l=1}^{n}\sum_{c=1}^{n}\mathcal{Y}\log\mathcal{P}^{o}$ and $\mathcal{L}^{ar}_{\theta_s} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{n}\left\|p_j - p^a_j\right\|_2^2$ for patch order distortion and patch appearance distortion in the student branch. Where n is the number of patches for each image, $\mathcal{Y}$ represent the order ground truth and $\mathcal{P}^o$ represent the network's patch order prediction, p_j and p^a_j represent image original appearance and reconstruction prediction.

Finally, the total loss is defined in Eq. 3, where $\mathcal{L}^{oc}_{\theta_s}$ is patch order classification loss, $\mathcal{L}^{ar}_{\theta_s}$ is patch appearance restoration loss, $\mathcal{L}^{G}_{\theta_s,\theta_t}$ is the global patch embedding consistency loss and $\mathcal{L}^{L}_{\theta_s,\theta_t}$ is the local patch embedding consistency loss. $\mathcal{L}^{oc}_{\theta_s}$ and $\mathcal{L}^{ar}_{\theta_s}$ empower the model to learn high-level anatomical structures. The $\mathcal{L}^{G}_{\theta_s,\theta_t}$ equips the model to learn the coarse-grained granularity and synthetical anatomy from global patch embeddings. $\mathcal{L}^{L}_{\theta_s,\theta_t}$ lets the model learn fine-grained and precise anatomical structures from local patch embeddings of overlapped parts. $\overline{y_s}$

$$\mathcal{L} = \mathcal{L}^{oc}_{\theta_s} + \mathcal{L}^{ar}_{\theta_s} + \mathcal{L}^{G}_{\theta_s,\theta_t} + \mathcal{L}^{L}_{\theta_s,\theta_t} \tag{3}$$

4 Experiments

4.1 Datasets and Implementation Details

Pretraining Settings. We pretrain PEAC with Swin-B as the backbone on unlabeled ChestX-ray14 [28] dataset. Our PEAC and PEAC^{-1} models utilize Swin-B as the backbone, pre-trained on an image size of 448×448 and fine-tuned on 448×448 and 224×224 respectively. PEAC^{-3} adopts ViT-B as the backbone, pre-trained and fine-tuned on an image size of 224×224. As for the prediction heads in the student branch, we use two single linear layers for the classification (patch order) and restoration tasks (patch appearance), and two 3-layer MLP for the expanders of local and global features. The augments used in the student branch include 50% probability of patch appearance distortion [38] and 50% probability of shuffling patches. The weights of Student model are updated by the total loss while the Teacher model are updated by Exponential Moving Average (EMA) [26] after each iteration. More details are in the supplementary materials Section 1.1.

Target Tasks and Datasets. To assess the performance of PEAC models, we transfer them to four thoracic disease classification tasks ChestX-ray14 [28], CheXpert [16], NIH Shenzhen CXR [17], RSNA Pneumonia [1] and one chest organ segmentation task JSRT [23]. For the segmentation task, we integrate Upernet [30] into the process. All downstream models share the same Swin-B backbone, where the encoder is initialized with PEAC pretrained weights and a final prediction decoder is re-initialized based on the number of classes for the target task. The total training rounds for classification and segmentation are 150 and 500 epochs, respectively. More details can be found in Section 1.2 of the supplementary materials.

4.2 Experimental Results

(1) PEAC outperforms fully-supervised pretrained models. We compare the performance of PEAC on four downstream tasks with different initialization methods. As shown in

Table 1, PEAC model outperforms both supervised ImageNet and ChestX-ray14 models, demonstrating that PEAC has learned transferable features for various medical image tasks. To show the statistical significance, we have conducted t-tests between PEAC/PEAC^{-1} and ImageNet-21K on all target tasks, yielding p-values: $(3.2\times10^{-9}, 4.0\times10^{-10}, 5.9\times10^{-5}, 0.018)/(2.9\times10^{-4}, 4.2\times10^{-8}, 0.033, 0.027)$ (all p-values$<$0.05), showing that the performance gains by PEAC, even by PEAC^{-1}, on all four target tasks are statistically significant at level of 0.05.

Table 1: PEAC models outperform fully-supervised pretrained models on ImageNet and ChestX-ray14 datasets in four target tasks across architectures. The best methods are bolded while the second best are underlined. We have conducted independent two sample t-text between the best vs. others and highlighted those in blue when they are not significantly different at $p = 0.05$ level. Transfer learning is inapplicable, when pretraining and target tasks are the same, and denoted by "–".

Backbone	Pretraining data	Pretraining method	ChestX-ray14	CheXpert	ShenZhen	RSNA Pneumonia
ResNet-50	No pretraining (i.e., training from scratch)		80.40 ± 0.05	86.60 ± 0.17	90.49 ± 1.16	70.00 ± 0.50
	ImageNet-1K	Fully-supervised	81.70 ± 0.15	87.17 ± 0.22	94.96 ± 1.19	73.04 ± 0.35
	ChestX-ray14	Fully-supervised	–	87.40 ± 0.26	96.32 ± 0.65	71.64 ± 0.37
ViT-B	No pretraining (i.e., training from scratch)		70.84 ± 0.19	80.78 ± 0.13	84.46 ± 1.65	66.59 ± 0.39
	ImageNet-21K	Fully-supervised	77.55 ± 1.82	83.32 ± 0.69	91.85 ± 3.40	71.50 ± 0.52
	ChestX-ray14	Fully-supervised	–	84.37 ± 0.42	91.23 ± 0.81	66.96 ± 0.24
	ChestX-ray14	PEAC^{-3} (self-supervised)	80.04 ± 0.20	88.10±0.29	96.69 ± 0.30	<u>73.77 ± 0.39</u>
Swin-B	No pretraining (i.e., training from scratch)		74.29 ± 0.41	85.78 ± 0.01	85.83 ± 3.68	70.02 ± 0.42
	ImageNet-21K	Fully-supervised	81.32 ± 0.19	87.94 ± 0.36	94.23 ± 0.81	73.15 ± 0.61
	ChestX-ray14	Fully-supervised	–	87.22 ± 0.22	91.35 ± 0.93	70.67 ± 0.18
	ChestX-ray14	PEAC^{-1} (self-supervised)	<u>81.90 ± 0.15</u>	<u>88.64 ± 0.19</u>	<u>97.17 ± 0.42</u>	73.70 ± 0.48
	ChestX-ray14	PEAC (self-supervised)	**82.78 ± 0.21**	**88.81 ± 0.57**	**97.39 ± 0.19**	**74.39 ± 0.66**

(2) PEAC outperforms self-supervised models pretrained on ImageNet. We compared PEAC models pretrained on ChestX-rays14 with the transformer-based models pretrained on ImageNet by various SOTA self-supervised methods including MoCo V3 [8], SimMIM [33], DINO [5], BEiT [2], and MAE [13]. The results in Table 4 show that our PEAC pretrained on moderately-sized unlabeled chestX-rays14 yields better results than the aforementioned SSL methods pretrained on larger ImageNet, revealing the transferability of PEAC via learning anatomical consistency with in-domain medical data.

Table 2: Comparing PEAC with SoTA baselines in terms of sensitivity to the number of training samples on ChestX-ray14. The baseline performance is adopted from DiRA [11].

Method	25%	50%	100%
MoCo-v2 [9, 11]	74.71	76.89	80.36
Barlow Twins [11, 36]	76.23	77.59	80.45
SimSiam [9, 11]	73.05	75.20	79.62
DiRA$_{\text{MoCo-v2}}$ [11]	<u>77.55</u>	<u>78.74</u>	<u>81.12</u>
PEAC	**77.78**	**79.29**	**82.78**

(3) PEAC outperforms recent self-supervised models pretrained on medical images. To demonstrate the effectiveness of representation learning via our proposed framework, we compare PEAC with SoTA CNN-based (SimSiam [6], MoCo V2 [7], Barlow Twins [36]) and transformer-based SSL methods (SimMIM [33]) pretrained on medical images as shown in Table 5. We also compare PEAC with these SoTA baselines in small data regimes shown in Table 2. The downstream dataset ChestX-ray14 is randomly seleted 25% and 50% when fintuning. Our method yields the best performance across four datasets and verifies its effectiveness on different backbones. We get several observations from the results: (1) In transformer backbones, our methods outperform SimMIM which demonstrates that patch order distortion and patch embedding consistency are effective; (2) In Swin-B backbone, our method outperforms POPAR^{-1} which shows the good generalization performance of patch

embedding consistency; (3) For different downstream label fractions, our method outperforming other baselines shows the robustness of patch embedding consistency.

(4) PEAC exhibits prominent transferability for segmentation tasks. As shown in Table 3, PEAC (1) surpasses SoTA SSL method POPAR [22] on two organ segmentation tasks (2) pretrained on the moderate-size Chest X-ray dataset yields competitive performance akin to fully supervised ImageNet models. These outcomes emphasize that employed as the backbone weights in the segmentation network, PEAC models facilitate accurate pixel-level predictions across three organ segmentation tasks. More details are shown in Fig. 8 of supplementary materials.

Table 3: Our PEAC leads the best or comparable performance on segmentation dataset JSRT over a self-supervised learning approach POPAR and fully supervised pretrained model.

Pretrain Method	Pretrain Dataset	JSRT Lung	JSRT Heart	JSRT Clavicle
Scratch	None	97.48 ± 0.08	92.87 ± 1.31	87.52 ± 1.78
Supervised	ImageNet-1K	**97.99 ± 0.01**	95.12 ± 0.08	92.18 ± 0.02
POPAR	ChestX-ray14	97.88 ± 0.12	93.67 ± 0.95	91.78 ± 0.53
PEAC	ChestX-ray14	97.97 ± 0.01	**95.21 ± 0.02**	**92.19 ± 0.04**

Table 4: Even downgraded PEAC^{-1} and PEAC^{-3} outperform SoTA self-supervised ImageNet in four target tasks. The best results are bolded and the second best are underlined.

Backbone	Pretrained dataset	Method	ChestX-ray14	CheXpert	ShenZhen	RSNA Pneumonia
ViT-B	ImageNet	MoCo V3	79.20 ± 0.29	86.91 ± 0.77	85.71 ± 1.41	72.79 ± 0.52
		SimMIM	79.55 ± 0.56	87.83 ± 0.46	92.74 ± 0.92	72.08 ± 0.47
		DINO	78.37 ± 0.47	86.91 ± 0.44	87.83 ± 7.20	71.27 ± 0.45
		BEiT	74.69 ± 0.29	85.81 ± 1.00	92.95 ± 1.25	72.78 ± 0.37
		MAE	78.97 ± 0.65	87.12 ± 0.54	93.58 ± 1.18	72.85 ± 0.50
	ChestX-ray14	PEAC^{-3}	80.04 ± 0.20	88.10 ± 0.29	96.69 ± 0.30	73.77 ± 0.39
Swin-B	ImageNet	SimMIM	81.39 ± 0.18	87.50 ± 0.23	87.86 ± 4.92	73.15 ± 0.73
	ChestX-ray14	PEAC^{-1}	81.90 ± 0.15	88.64 ± 0.19	97.17 ± 0.42	73.70 ± 0.48
		PEAC	**82.78 ± 0.21**	**88.81 ± 0.57**	**97.39 ± 0.19**	**74.39 ± 0.66**

Table 5: To speed up the training process, we compare the performance of downstream tasks using image resolution of 224. All models are pretrained on the ChestX-ray14 dataset.

Backbone	Pretrained dataset	Method	ChestX-ray14	CheXpert	ShenZhen	RSNA Pneumonia
ResNet-50	ChestX-ray14	SimSiam	79.62 ± 0.34	83.82 ± 0.94	93.13 ± 1.36	71.20 ± 0.60
		MoCoV2	80.36 ± 0.26	86.42 ± 0.42	92.59 ± 1.79	71.98 ± 0.82
		Barlow Twins	80.45 ± 0.29	86.90 ± 0.62	92.17 ± 1.54	71.45 ± 0.82
ViT-B	ChestX-ray14	SimMIM	79.20 ± 0.19	83.48 ± 2.43	93.77 ± 1.01	71.66 ± 0.75
		PEAC^{-3}	80.04 ± 0.20	88.10 ± 0.29	96.69 ± 0.30	**73.77 ± 0.39**
Swin-B	ChestX-ray14	SimMIM	79.09 ± 0.57	86.75 ± 0.96	93.03 ± 0.48	71.99 ± 0.55
		POPAR^{-1}	80.51 ± 0.15	88.16 ± 0.66	96.81 ± 0.40	73.58 ± 0.18
		PEAC^{-1}	**81.90 ± 0.15**	**88.64 ± 0.19**	**97.17 ± 0.42**	73.70 ± 0.48

4.3 Ablations

(1) Transformer grid-based matching is more stable. Calculating local consistency requires the correspondence between patches across views. Existing block matching [4, 35] is derived based on the adjacent position or cosine similarity of embedding vectors between blocks, leading to inexact matches, resulting in training instability and unreliability. By contrast, our grid-based matching (illustrated in Fig. 2) is more stable as shown in Fig. 4.

(2) Local consistency improves performance. We remove the local loss and use several different backbones to demonstrate the effectiveness of local embedding consistency. The discussions and table of the results can be found in the supplementary materials Section 3.1.

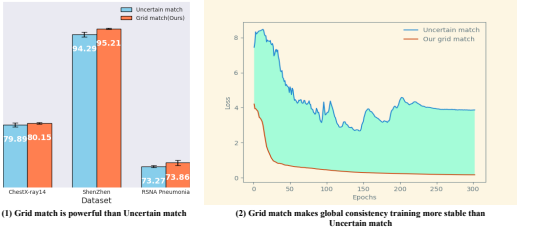


Figure 4: (1) Grid-based matching show better performance than block matching in VICRegL's [4] uncertain match. (2) The cures for the global-consistency loss show more stability when training with grid-based match.

(3) Student-teacher model (global consistency) boosts all one branch methods. In our experiments we found that the student-teacher model can be applied to one-branch self-supervised methods and boosts their performance. The results are included in the supplementary materials Section 3.2.

4.4 Visualization of Upstream Models

(1) Cross-patient and cross-view correspondence. To show that our PEAC can learn a variety of anatomical structure effectively, we match patch-level features across patients and different views of the same image, as shown in Fig. 5. Our observations suggest that these features capture semantic regions and exhibit robustness across samples with large morphological differences and both sexes. Similar semantic regions are also stably captured in different views of the same sample. We also compared our results to models trained with other methods, which exhibited obvious incorrect matches. Further details are available in the supplementary materials Section 4.1.

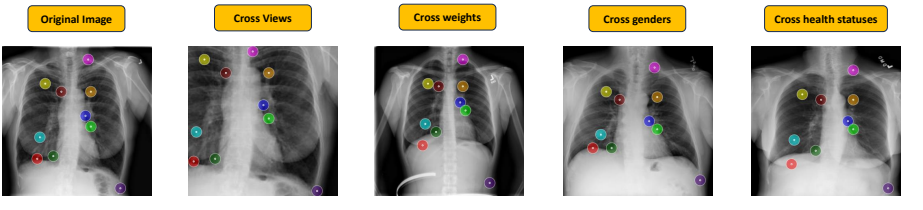


Figure 5: Our PEAC model can effectively localize arbitrary anatomical structures across views of the same patient and across patients of different genders and weights and of health and disease.

(2) t-SNE of landmark anatomies across patients. We use t-SNE on 7 labeled local landmark anatomies across 1000 patient images. Each local anatomy is labeled with different color which corresponds with the t-sne cluster color in Fig. 6. This shows our PEAC's ability to learn a valid embedding space among different local anatomical structures.

(3) Zero-shot co-segmentation. Without finetuning on downstream tasks, we jointly segment analogous structures common to all images in a given set as shown in Fig. 7. In our segmentation results, different anatomical structures clearly segmented as common features suggests that our proposed model demonstrates proficiency in extracting and representing the distinguishing features of various anatomical regions.

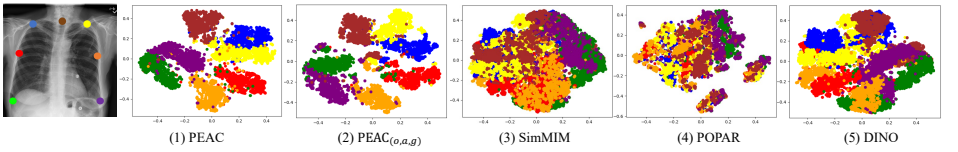


Figure 6: We use t-SNE to contrast our PEAC's landmark embeddings with SimMIM, POPAR, and DINO. PEAC clearly delineates each landmark embedding, demonstrating valid embedding space. However, the reduced PEAC$_{(o,a,g)}$, lacking local consistency, shows intermingled landmarks, emphasizing local consistency's role in defining distinct anatomical structure embeddings.

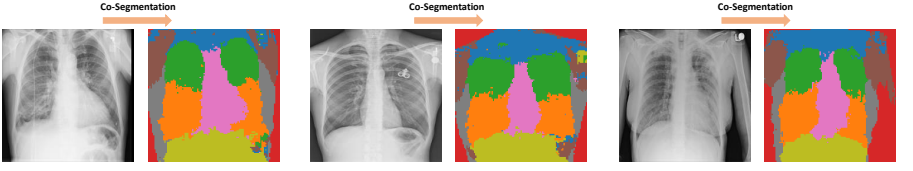


Figure 7: We semantically co-segment common structure of images in a zero-shot scenario. The cervico scapular region, upper lobe of lungs, lower lobe of lungs, mediastinum, and abdominal cavity are clearly segmented as common features.

5 Conclusion

We propose a novel self-supervised learning approach, denoted as PEAC, designed to enhance the consistency in learning visual representations of anatomical structures within medical images. The vital technique in PEAC is our novel yet reliable grid-based matching which guarantees both global and local consistency in anatomy. Through extensive experiments, we demonstrate the effectiveness of our scheme. By accurately identifying the features of each common region across patients of different genders and weights and across different views of the same patients, PEAC exhibits a heightened potential for enhanced AI in medical image analysis.

6 Acknowledgement

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] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. RSNA pneumonia detection challenge (2018).

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, 2022.

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *arXiv preprint arXiv:2210.01571*, 2022.

[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*, pages 9650–9660, 2021.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[8] 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*, pages 9640–9649, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[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*, pages 20824–20834, 2022.

[12] 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*, pages 9729–9738, 2020.

[13] 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*, pages 16000–16009, 2022.

[14] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[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):475, 2014.

[18] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[19] Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

[20] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[22] 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 *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 77–87. Springer, 2022.

[23] 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: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[24] Nima Tajbakhsh, Holger Roth, Demetri Terzopoulos, and Jianming Liang. Guest Editorial Annotation-Efficient Deep Learning: The Holy Grail of Medical Imaging. *IEEE Transactions on Medical Imaging*, 40(10):2526–2533, oct 2021. ISSN 1558254X. doi: 10.1109/TMI.2021.3089292.

[25] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings*, pages 661–673. Springer, 2021.

[26] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[28] 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.

[29] 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*, pages 3024–3033, 2021.

[30] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[31] 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*, pages 10539–10548, 2021.

[32] 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*, pages 16684–16693, 2021.

[33] 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*, pages 9653–9663, 2022.

[34] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8354–8363, 2022.

[36] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[37] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[38] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.