Learning Anatomically Consistent Embedding for Chest Radiography

Supplementary Materials

1 Experiments

1.1 Pretraining settings

We implement PEAC on ViT-B [5] and Swin-B [8] for their notable scalability, global receptibility, and interpretability [4, 5]. Both PEACs are trained on ChestX-ray14 [14] by amalgamating the official training and validation splits. In PEAC ViT-B, input images of size 224×224 lead to 196 (14×14) shufflable patches, while in PEAC Swin-B, it results in 49 (7×7) shufflable patches due to the Swin hierarchical architecture. To learn the same contextual relationship as in PEAC ViT-B, we pretrain PEAC Swin-B with images of size 448×448, but the tissue (physical) size covered by the images remains unchanged, resulting in the same 196 (14×14) shufflable patches in terms of the (physical) tissue size.

In PEAC, a multi-class linear layer is designated for patch order classification (Eq. 1), and a single convolutional block is employed for patch appearance restoration (Eq. 2). The global and local consistency branches utilize two 3-layer MLPs as expanders before computing consistency losses. When training PEAC, we use a learning rate of 0.1, a momentum of 0.9 for the SGD optimizer, a warmup period of 5 epochs, and a batch size of 8. The teacher model is updated after each iteration via EMA with an updating parameter of 0.999. We utilize four Nvidia RTX3090 GPUs for training PEAC models with images of size 224×224 for 300 epochs, but we reduce the number of epoch to 150 when the image size is 448×448.

1.2 Target Tasks and Datasets

We evaluate our PEAC models by finetuning on four classification target tasks ChestX-ray14 [14], CheXpert [6], NIH Shenzhen CXR [7], RSNA Pneumonia [1] and one segmentation task JSRT [13]:

- **ChestX-ray14** [14], which contains 112K frontal-view X-ray images of 30805 unique patients with the text-mined fourteen disease image labels (where each image can have multi-labels). We use the official training set 86K (90% for training and 10% for validation) and testing set 25K.

$*$ Equal contribution. † Corresponding author.

- **CheXpert** [6], which includes 224K chest radiographs of 65240 patients and capturing 14 thoracic diseases. We use the official training data split 224K for training and validation set 234 images for testing. We train on the 14 thoracic diseases but follow the standard practice by testing on 5 diseases (Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion).

- **NIH Shenzhen CXR** [7], which contains 326 normal and 336 Tuberculosis (TB) frontal-view chest X-ray images. We split 70% of the dataset for training, 10% for validation and 20% for testing which are the same with [10];

- **RSNA Pneumonia** [1], which consists of 26.7K frontal view chest X-ray images and each image is labeled with a distinct diagnosis, such as Normal, Lung Opacity and Not Normal (other diseases). 80% of the images are used to train, 10% to valid and 10% to test. These target datasets are composed of both multi-label and multi-class classification tasks with various diseases.

- **JSRT** [13], which is a organ segmentation dataset including 247 frontal view chest X-ray images. All of them are in 2048×2048 resolution with 12-bit gray-scale levels. Both lung, heart and clavicle segmentation masks are available for this dataset. We split 173 images for training, 25 for validation and 49 for testing.

1.3 Finetuning Setting

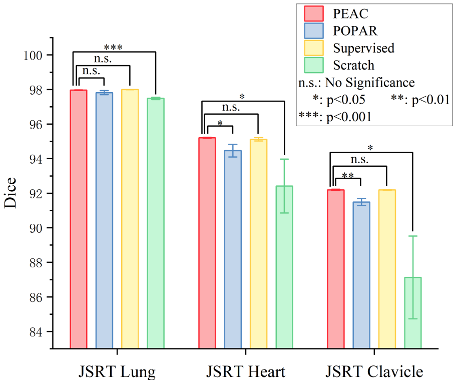

Figure 8: Comparison of segmentation results on JSRT dataset. To investigate the performance on segmentation tasks, we compare PEAC with SSL method POPAR, fully-supervised model, and model training from scratch.

We transfer the PEAC pretrained models to each target task by fine-tuning the whole parameters for the target classification tasks. For the target classification tasks, we concatenate a randomly initialized linear layer to the output of the classification (CLS) token of PEAC ViT-B models. Due to the structural difference with ViT-B model, PEAC Swin-B models don't equip the CLS token and we add an average pooling to the last-layer feature maps, then feed the feature to the randomly initialized linear layer. The AUC (area under the ROC curve) is used to evaluate the multi-label classification performance (ChestX-ray14, CheXpert and NIH Shenzhen CXR), while the accuracy is used to assess the multi-class classification performance (RSNA Pneumonia). For the target segmentation task, we use UperNet [15] as the training model. We concatenate pretrained Swin-B and randomly initialized prediction head for segmenting. In JSRT dataset, we independently train 3 models for the three organs lung, heart and clavicle and the Dice is used to evaluate the segmentation performance, as shown in and Fig. 8 . Following [8], in fine-tuning experiments we use AdamW [9] optimizer with a cosine learning rate scheduler, linear warm up of 20 epochs while the overall epoch is 150, and 0.0005 for the maximum learning rate value. The batch sizes are 32 and 128 for image sizes of 448 and 224, respectively. We train with single Nvidia RTX3090 24G GPU for performing each experiment.

Table 6: We add the four loss functions one by one to show the effectiveness of our method in terms of performance. All models in the ablation studies are pretrained on ChestX-ray14 [14] with Swin-B backbone at two different image resolutions, and they are also fine-tuned at two different image resolutions as denoted by PT→FT in the table. Our official implementation PEAC achieves the best performance or the second best on three target tasks with pretraining and finetuning resolutions set at 448 × 448.

PEAC Version	Shuffled patches	PT→FT	Transformations		POPAR Losses		PEAC Losses		Target Tasks		
			OD	AD	$\mathcal{L}^{oc}_{\theta_s}$	$\mathcal{L}^{ar}_{\theta_s}$	$\mathcal{L}^{G}_{\theta_s,\theta_t}$	$\mathcal{L}^{L}_{\theta_s,\theta_t}$	ChestX-ray14	ShenZhen	RSNA Pneumonia
$\text{PEAC}^{-2}_{(o)}$	49	$224^2 \rightarrow 224^2$	✓	✗	✓	✗	✗	✗	78.58±0.17	92.65±0.65	71.46±0.41
$\text{PEAC}^{-2}_{(a)}$			✗	✓	✗	✓	✗	✗	79.35±0.18	93.85±0.09	72.38±0.15
$\text{PEAC}^{-2}_{(o,a)}$			✓	✓	✓	✓	✗	✗	79.57±0.22	95.10±0.20	72.59±0.13
$\text{PEAC}^{-2}_{(g)}$	49	$224^2 \rightarrow 224^2$	✗	✗	✗	✗	✓	✗	80.85±0.14	96.59±0.11	73.42±0.41
$\text{PEAC}^{-2}_{(o,g)}$			✓	✗	✓	✗	✓	✗	81.13±0.18	96.70±0.11	73.75±0.04
$\text{PEAC}^{-2}_{(o,g,l)}$			✓	✗	✓	✗	✓	✓	81.09±0.35	97.00±0.28	**74.42±0.34**
$\text{PEAC}^{-2}_{(o,a,g)}$			✓	✓	✓	✓	✓	✗	81.25±0.16	96.91±0.07	73.35±0.19
PEAC^{-2}			✓	✓	✓	✓	✓	✓	81.38±0.03	97.14±0.10	74.19±0.15
$\text{PEAC}^{-1}_{(o,a,g)}$	196	$448^2 \rightarrow 224^2$	✓	✓	✓	✓	✓	✗	81.51 ± 0.22	97.07±0.37	73.63 ± 0.42
PEAC^{-1}			✓	✓	✓	✓	✓	✓	81.90±0.15	97.17±0.42	73.70±0.48
$\text{PEAC}_{(o,a,g)}$	196	$448^2 \rightarrow 448^2$	✓	✓	✓	✓	✓	✗	82.67±0.11	97.15±0.40	74.18±0.52
PEAC			✓	✓	✓	✓	✓	✓	**82.78±0.21**	**97.39±0.19**	74.39±0.66

Table 7: The local consistency loss in PEAC consistently improves the performance across methods and target tasks.

Method	Backbone	Transformations		POPAR Losses		PEAC Losses		Target Tasks		
		OD	AD	$\mathcal{L}^{oc}_{\theta_s}$	$\mathcal{L}^{ar}_{\theta_s}$	$\mathcal{L}^{G}_{\theta_s,\theta_t}$	$\mathcal{L}^{L}_{\theta_s,\theta_t}$	ChestX-ray14	ShenZhen	RSNA Pneumonia
VICRegL	ConvNeXt-B	✗	✗	✗	✗	✗	✗	79.89±0.34	94.29±0.40	73.27±0.15
$\text{VICRegL}_{(l)}$		✗	✗	✗	✗	✗	✓	80.15±0.11	95.21±0.11	73.86±0.43
SimMIM	Swin-B	✗	✗	✗	✗	✗	✗	79.09±0.57	93.03±0.48	71.99±0.55
$\text{SimMIM}_{(g)}$		✗	✗	✗	✗	✓	✗	81.42± 0.04	97.11± 0.26	73.95± 0.18
$\text{SimMIM}_{(g,l)}$		✗	✗	✗	✗	✓	✓	81.67± 0.04	97.86± 0.07	74.25±0.24
$\text{PEAC}_{(o,a,g)}$	Swin-B	✓	✓	✓	✓	✓	✗	82.67±0.11	97.15±0.40	74.18±0.52
PEAC		✓	✓	✓	✓	✓	✓	**82.78±0.21**	**97.39±0.19**	**74.39±0.66**

2 Ablation Studies: PEAC versions and their performance

Our PEAC involves four losses:

- Patch order classification loss defined in the main paper as

$$\mathcal{L}^{oc}_{\theta_s} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{l=1}^{n}\sum_{c=1}^{n}\mathcal{Y}\log\mathcal{P}^{o} \tag{1}$$

- Patch appearance restoration loss defined in the main paper as

$$\mathcal{L}_{\theta_s}^{ar} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{n} \|p_j - p_j^a\|_2^2 \tag{2}$$

- Global patch embedding consistency loss defined in the main paper as $\mathcal{L}_{\theta_s,\theta_t}^{G} = \mathcal{L}_{\theta_s,\theta_t}^{global} + \widetilde{\mathcal{L}}_{\theta_s,\theta_t}^{global}$. $\mathcal{L}_{\theta_s,\theta_t}^{global}$ and $\widetilde{\mathcal{L}}_{\theta_s,\theta_t}^{global}$ are computed from an exchanged inputs into the Student and Teacher models.

$$\mathcal{L}_{\theta_s,\theta_t}^{global} \triangleq \|\overline{y_s} - \overline{y_t}\|_2^2 = 2 - 2\cdot\frac{y_s - y_t}{\|y_s\|_2\cdot\|y_t\|_2} \tag{3}$$

- Local patch embedding consistency loss defined in the main paper as $\mathcal{L}_{\theta_s,\theta_t}^{L} = \mathcal{L}_{\theta_s,\theta_t}^{local} + \widetilde{\mathcal{L}}_{\theta_s,\theta_t}^{local}$. $\mathcal{L}_{\theta_s,\theta_t}^{local}$ and $\widetilde{\mathcal{L}}_{\theta_s,\theta_t}^{local}$ are computed from an exchanged inputs into the Student and Teacher models.

$$\mathcal{L}_{\theta_s,\theta_t}^{local} \triangleq \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\cdot(\sum_{i=1}^{z}\|\overline{p_{m_i}} - \overline{p_{n_i}}\|_2^2) \tag{4}$$

We remove some ingredients from our official implementation PEAC and the results (Table 6) show the effectiveness of all loss functions. The POPAR versions involve OD (patch order distortion) and AD (patch appearance distortion) which are studied in [12] and the losses include patch order classification loss $\mathcal{L}_{\theta_s}^{oc}$, patch appearance restoration loss $\mathcal{L}_{\theta_s}^{ar}$. The downgraded version PEAC$_{(o)}^{-2}$ only include OD, in this circumstance we only compute $\mathcal{L}_{\theta_s}^{oc}$ and neglect the $\mathcal{L}_{\theta_s}^{ar}$. Correspondingly, only AD is added for the downgraded version PEAC$_{(a)}^{-2}$, in this case we only compute $\mathcal{L}_{\theta_s}^{ar}$ and neglect the $\mathcal{L}_{\theta_s}^{oc}$. The PEAC versions involve the four loss functions mentioned above. Under the same settings (the same shuffled patches and the same pretraining and fine-tuning resolutions), we added these loss functions one by one, and the downstream tasks performance improve successively shown in Table 6.

Our pretraining and fine-tuning setting include two resolutions 448×448 and 224×224. The downgraded versions PEAC^{-2} contain 49 pretraining shuffled patches and are pretrained and fine-tuned on 224 size of images while the downgraded versions PEAC^{-1} include 196 shuffled patches and are pretrained on 448 and fine-tuned on 224 size of images. And the performances on our official implementation PEAC (pretrained and fine-tuned on 448 images) are the best. To accelerate the training process, we only pretrain two versions PEAC and PEAC$_{(o,a,g)}$ on 448 images.

Table 8: The global loss in PEAC consistently boosts the performance across methods and target tasks.

Method	Transformations		POPAR Losses		PEAC Losses		Target Tasks		
	OD	AD	$\mathcal{L}_{\theta_s}^{oc}$	$\mathcal{L}_{\theta_s}^{ar}$	$\mathcal{L}_{\theta_s,\theta_t}^{G}$	$\mathcal{L}_{\theta_s,\theta_t}^{L}$	ChestX-ray14	ShenZhen	RSNA Pneumonia
SimMIM	✗	✗	✗	✗	✗	✗	79.09±0.57	93.03±0.48	71.99±0.55
SimMIM$_{(g)}$	✗	✗	✗	✗	✓	✗	**81.42±0.04**	**97.11±0.26**	73.95±0.18
POPAR$_{od}^{-2}$	✓	✗	✓	✗	✗	✗	78.58±0.17	92.65±0.65	71.46±0.41
PEAC$_{(o,g)}^{-2}$	✓	✗	✓	✗	✓	✗	81.13±0.18	96.70±0.11	73.75±0.04
POPAR^{-2}	✓	✓	✓	✓	✗	✗	79.57±0.22	95.10±0.20	72.59±0.13
PEAC$_{(o,a,g)}^{-2}$	✓	✓	✓	✓	✓	✗	81.38±0.03	96.91±0.10	**74.19±0.15**

3 Ablations: Local and global consistency

3.1 PEAC local consistency improves performance

We add the local consistency loss based on several methods VICRegL [3], SimMIM [16] shown in Table 7. In the instance of VICRegL, ConvNeXt serves as the backbone, with the subsequent addition of local consistency loss precipitating notable enhancements in performance across all three target tasks. The SimMIM methodology employs Swin-B as its backbone, with the sequential addition of global and local consistency losses leading to marked improvements in performance. Moreover, the removal of local consistency loss from our PEAC method corresponds to a decline in performance across the target classification tasks. This evidence underscores the efficacy of our proposed grid-matched local consistency loss.

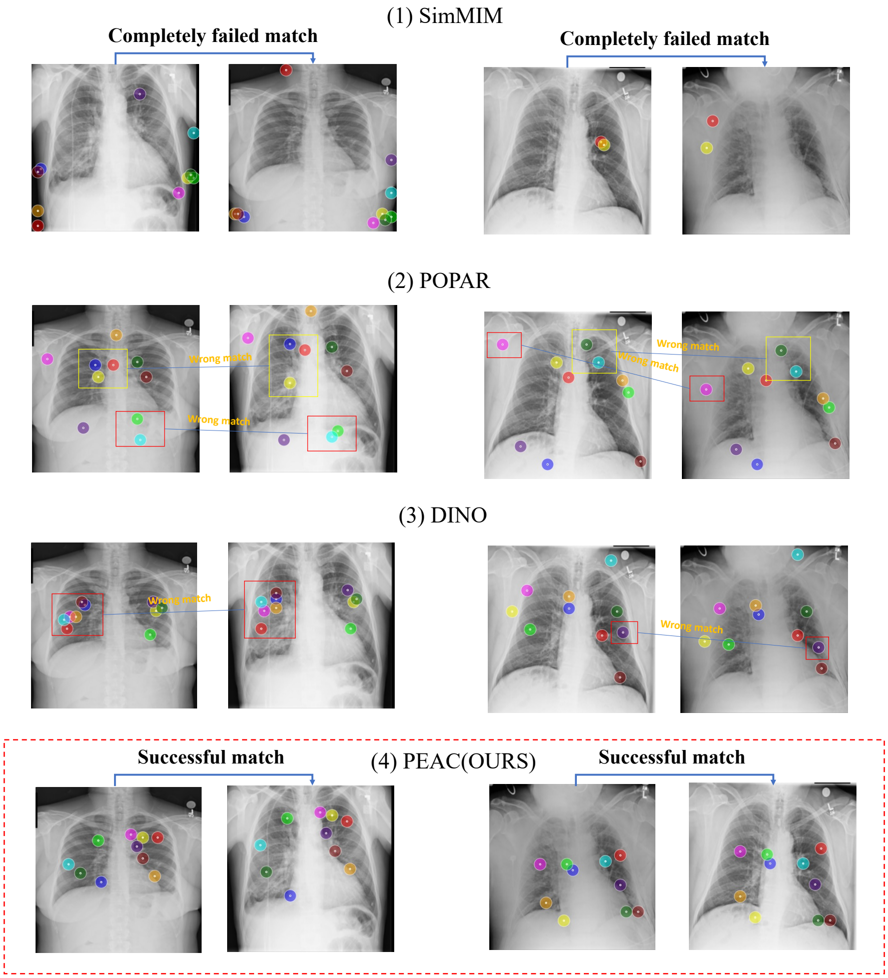

Figure 9: Comparing PEAC with DINO, POPAR, and SimMIM in matching anatomical strictures across distinct patients. For the same pair of patient images, our PEAC provides the most reliable anatomy matching.

3.2 PEAC global consistency boosts performance

Corresponding to the main paper in Section 4.3 (3) our experiments in Table 8 demonstrate that using Teacher-Student model with global embedding consistency can boost one branch methods. We conduct experiments based on SimMIM and our own method which are all based on Swin-B backbone, pretrained on ChestX-ray14 [14], pretrained and fine-tuned on 224 image resolution. When adding teacher branch for SimMIM to compute the global embedding consistency loss, the classification performances of SimMIM$_{(g)}$ for the three target tasks are significantly improved. Importantly, the input images of the two branches are the two global views which are grid-wise cropped using our method and the student branch in SimMIM$_{(g)}$ gets the masked patches as SimMIM while the teacher branch gets no augmentations for the input images. We also add teacher branch to the one branch methods POPAR$_{od}^{-2}$ and POPAR^{-2} for computing the global consistency loss. The downstream performances on the two branches Teacher-Student models PEAC$_{(o,g)}^{-2}$ and PEAC$_{(o,a,g)}^{-2}$ are much better than one branch methods POPAR$_{od}^{-2}$ and POPAR^{-2}.

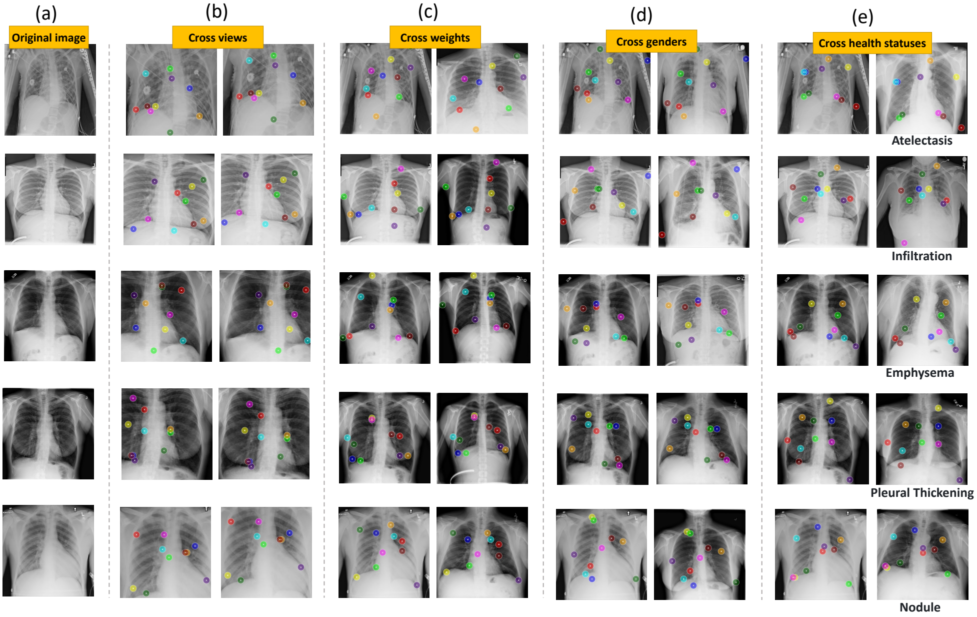

Figure 10: Establishing anatomical correspondence across views, across subject weights, across genders, and across health statuses. (a) original images of patients with no detected pulmonary disease. (b) cross-view correspondences, utilizing cropped images from the original set. (c) cross-weight correspondences among patients with significant differences in weight. (d) cross-gender correspondences between male and female patients. (e) correspondences among patients with varied health conditions. The left chest X-rays in (c), (d), and (e) are the same original images from (a), contrasting with the right ones that feature patients with distinct genders, weights, and health statuses. In (e), we also indicate diseases under the respective images.

4 Visualization of Upstream Models

4.1 Cross-patient and cross-view correspondence

To investigate the promotion of our method for sensing local anatomy, we match small local patches across two patients' and one patient's different views of X-ray. Fig. 9 shows the cross-patient correspondence of our PEAC and other methods. Following [2] we divide each image with a resolution of 224 into 196 image patches using ViT-B backbone, and match the patch embedding of each image patch to the most similar patch embedding in another image. Finally, we selected the top 10 most similar image patches with K-means and drew the correspondence points. By comparing the correspondence results of our methods with SimMIM, POPAR and DINO in Fig. 9, we learn that our method PEAC can learn the local anatomy more precisely. The details of the algorithm are shown in Fig. 11.

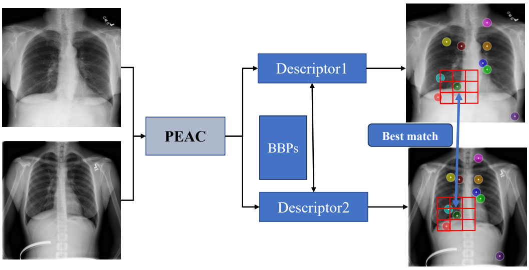

Figure 11: Pipeline of Corresponding patch match. After training, during inference, PEAC has only one single image as input. We eatablish cross-view or cross-patient correspondences by first computing dense embedding for each image. Specifically, we Generate 53×53 dense embeddings using a 14x14 grid with patch size of 16×16. The grid is shifted four times by 4 pixels to the right and for each of the shift to the right, it is further shifted four times by 4 pixels downward, leading to a grid of 53×53. Then we calculate the cosine similarity metric of embedding vectors from one image to another and find N Best-Buddies Pairs (BBPs) [11]. Finally, from the BBPs we selected the top10 most similar image patch pairs with K-means and drew the correspondence points.

We also use our PEAC method to match anatomical structures from a patient with no finding (disease) to patients of different weights, different genders, and different health statuses as shown in Fig. 10. The results show that our PEAC can consistently and precisely capture similar anatomies across different views of the same patients and across patients of opposite genders, different weights, and various health statuses.

References

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

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

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

[4] 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.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] 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.

[7] 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.

[8] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[10] DongAo Ma, Mohammad Reza Hosseinzadeh Taher, Jiaxuan Pang, Nahid UI Islam, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Benchmarking and boosting transformers for medical image classification. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 12–22. Springer, 2022.

[11] Shaul Oron, Tali Dekel, Tianfan Xue, William T Freeman, and Shai Avidan. Best-buddies similarity—robust template matching using mutual nearest neighbors. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1799–1813, 2017.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.