ACE: Anatomically Consistent Embeddings in Composition and Decomposition

Ziyu Zhou[*1,2] Haozhe Luo[*2,3] Mohammad Reza Hosseinzadeh Taher[2] Jiaxuan Pang[2]
Xiaowei Ding[†1] Michael Gotway[4] Jianming Liang[†2]
[1]Shanghai Jiao Tong University [2]Arizona State University [3]University of Bern [4]Mayo Clinic

A. Implementation Details

A.1. Pseudo code implementation

We illustrate the method of ACE in the main paper and propose a pseudo-code implementation of local consistency in the Algorithm. 1.

Algorithm 1 local consistency Pytorch pseudo-code.

```
# fs, ft: student and teacher encoders
# Cs, Ds: composer and decomposer head
# O1, O2: overlap area mask (0/1) of crop C1 and C2
# kernel: gaussian kernel to smooth the matrix target
ft.params = fs.params
ft.requires_grad = False
for C1, C2, O1, O2 in loader: # load a minibatch
    C1, C2 = augment(C1), augment(C2) # random views

    s1, s2 = fs(C1), fs(C2) # student output
    t1, t2 = ft(C1), ft(C2) # teacher output

    # compute composition loss
    s1 = Cs(s1) # input composer head
    O1 = maxpool(O1) # pool size 2x2
    loss_comp = ComputeLoss(O1, O2, s1, t2)

    # compute decomposition loss
    s2 = Ds(s2) # input composer head
    O2 = interpolate(O2) # upsample the overlap mask
    loss_decomp = ComputeLoss(O1, O2, s1, t2)

    loss = loss_comp/2 + loss_decomp/2
    loss.backward() # back-propagate

    # student, teacher updates
    update(fs) # Adam
    ft.params = m*gt.params + (1-m)*gs.params

def ComputeLoss(O1, O2, s, t):
    # compute matching matrix
    M = torch.mul(t.flattern(), s.flattern())

    # compute matrix target
    T = torch.zeros(len(t.flattern()),len(s.flattern()))
    idx1 = torch.nonzero(O1)
    idx2 = torch.nonzero(O2)

    # apply gaussian weights on the target coordinates
    T[idx2, idx1] = kernel

    return - (T * log(M)).sum(dim=1).mean()
```

A.2. Pretraining and testing datasets

We evaluate our ACE on chest X-rays and fundus photography, pretraining on ChestX-ray14 [18] and EyePACS [4] datasets respectively. The pretrained ACE models are validated on target tasks including the following datasets:

- **ChestX-ray14** [18], 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. The downstream models are trained to predict 14 pathologies in a multi-label classification setting and the mean AUC score is utilized to evaluate the classification performance. In addition to image-level labeling, the datasets provides bounding box annotations for 880 images in test set. Of this set of images, bounding box annotations are available for 8 out of 14 thorax diseases. After finetuning, we use the bounding box annotations in test set to assess the accuracy of pathology localization in a weakly-supervised setting. Besides, we compile a dataset of 1,000 images from test set, each annotated by experts with distinct anatomical landmarks. We use these labeled landmarks for anatomical embeddings analysis (see main paper Sec. 5.2-1 and Sec. B.1), unsupervised key-point correspondence (see main paper Sec. 5.2-2), key-point detection (see Sec. B.2)
- **NIH Shenzhen CXR** [9], 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 [13];
- **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.
- **JSRT** [16], which is an 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. The heart and clavicle segmentation masks are utilized for this dataset. We split 173 images for training, 25 for validation and 49 for testing.
- **ChestX-Det** [10], which is a disease segmentation dataset and an improved version of ChestX-

∗ Equal contribution. † Corresponding author.

Det10 [11]. This dataset contains 3,578 images with instance-level annotations for 13 common thoracic pathology categories, sourced from the NIH ChestX-ray14 dataset. Annotations were provided by three board-certified radiologists, and the dataset includes additional segmentation annotations. We consolidated all the diseases into one region and the goal of segmenting this dataset is to distinguish between diseased and non-diseased areas for each image. There are official split for training and testing sets and we split 10% images from training set for validation.

- **SIIM-ACR** [2], a dataset resulting from a collaboration between SIIM, ACR, STR, and MD.ai, contains 12,089 chest X-ray images. It is the largest public pneumothorax segmentation dataset to date, comprising 3,576 pneumothorax images and 9,420 non-pneumothorax images, all of which are available in 1024x1024 pixel resolution. We randomly divided the dataset into training (80%), validation (10%) and testing (10%). The segmentation performance is measured by the mean Dice which average the dice of pneumothorax non-pneumothorax images.
- **EyePACS** [4], a diabetic retinopathy (DR) classification dataset for identifying signs of diabetic retinopathy in eye images. The clinician has rated the presence of diabetic retinopathy in each image on a scale of 0 to 4, 0 for no DR, 1 for mild DR, 2 for moderate DR, 3 for severe DR and 4 for proliferative DR. There are 53,576 unlabeled images and 35,126 with labels. We randomly split the labeled images into training (80%), validation (10%) and testing (10%) set for downstream evaluation, and we merge the training, validation and unlabeled sets for pretraining.
- **FIRE** [6], the dataset comprises 134 pairs of images obtained from 39 patients, with each pair annotated with specific corresponding keypoints. In our target task, for each pair of images, one image is designated as the query image, and the task is to identify the corresponding anatomical structures in the key image. Additionally, we simultaneously visualize the predicted and ground truth keypoints in the key image.

A.3. Pretraining settings

We have trained two ACE models with Swin-B backbone using unlabeled images from ChestX-ray14 and EyePACS for the adaptation on chest X-ray and fundus imaging. Moreover, to generalize to other architecture we have trained ACE on ViT-B backbone on ChestX-ray14. Our ACE learning paradigm is similar to knowledge distillation [7], where a student network learns to match a teacher network's output. The weights of the student model θ_s are updated by back-propagation and the gradients of teacher model are stopped whose weights θ_t are updated using EMA (exponential moving average) from student. The update rule is $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$, where λ follows a cosine schedule from 0.996 to 1 during training.

The composer and decomposer heads are 2-layer MLPs to integrate and expand the local embeddings. In detail, the output of student or teacher encoder are patch embeddings with shape $14 \times 14 \times 1024$. Before the composer head, each $2 \times 2 \times 1024$ adjacent embeddings are concatenated and the patch embeddings are reshaped to $7 \times 7 \times 4096$, then they are input to a 2-layer MLP with input dimension 4096 and output dimension 1024 to get shape of $7 \times 7 \times 1024$ embeddings. Symmetrically, in the decomposer head, the $14 \times 14 \times 1024$ patch embeddings are input to a 2-layer MLP with input dimension 1024 and output dimension 4096 to expand the embeddings to $14 \times 14 \times 4096$, then each embedding is chunked into $2 \times 2 \times 1024$ and the output embeddings will be $28 \times 28 \times 1024$.

During the pretraining phase, we utilize a batch size of 8 images per GPU and train for a total of 100 epochs with 4 V100 (32G). The optimizer is AdamW and the initial learning rate is set to 5e-4 with a linear warm-up over the first 10 epochs. The weight decay starts at 0.04 and reaches 0.4 by the end of training, following a cosine schedule. The drop path rate is set to 0.1. Gradient clipping is applied with a maximum norm of 0.8 to ensure stable training dynamics.

A.4. Finetuning settings

For the target classification tasks, we concatenate a randomly initialized linear layer to the output of the classification (CLS) token of ViT-B pretrained models. For Swin-B pretrained models, we add an average pooling to the last-layer feature maps, then feed the feature to the randomly initialized linear layer. For the target segmentation task, we use UperNet [19] as the training model. We concatenate pretrained weights and randomly initialized prediction head for segmenting. Following [12], we employ the AdamW optimizer in conjunction with a cosine learning rate scheduler. We incorporate a linear warm-up phase spanning 20 epochs, within a total training duration of 150 epochs. The base learning rate is set at 0.0001. Each experiment is conducted using four V100 32 GPUs, with a batch size of 32 per GPU. For segmentation tasks, we retain the same setup and extend the training period to 500 epochs.

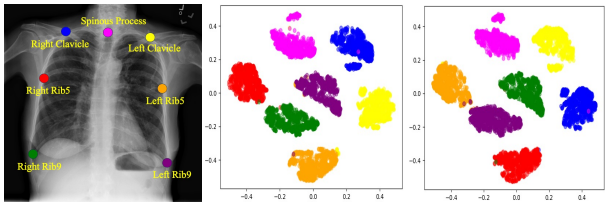
B. Additional Results

B.1. Emergent property: ACE understand anatomical symmetry.

Experimental Setup: We examine ACE's ability to capture the symmetry of anatomical structures in its learned embedding space. To do so, we consider $N = 7$ anatomical landmarks, including three pairs of mirrored structures and one structure located in the center of the chest, as shown

in Fig. 1–a. We extract size of 448^2 patches ($C = \{C_i\}_{i=1}^{N}$) around each landmark's location from the original images, and then use ACE's pretrained model to extract latent features for each landmark and its corresponding left and right flipped version ($\tilde{C} = \mathcal{T}(C)$). The extracted features of C and $\tilde{C}$ are visualized via t-SNE plots in Fig. 1–b and 1–c, respectively.

Results: As seen in Fig.1–b and Fig.1–c, ACE captures the symmetry of anatomical structures within its learned embedding space. For example, the right and left clavicles, which are visually symmetrical, are represented similarly in the embedding space. As seen, the blue cluster in Fig.1–b, corresponding to the right clavicle, closely matches the yellow cluster in Fig. 1–c, which represents the flipped left clavicle. A similar pattern is observed for other pairs, such as the left rib 5 and its flipped version, represented by the orange and red clusters in Fig.1–b and Fig.1–c, respectively. These observations demonstrate that ACE effectively captures the symmetry of anatomical structures in its learned embedding space as an *emergent* property.



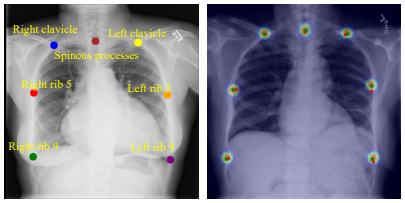
(a) (b) (c)

Figure 1. ACE reflects the symmetry of anatomical structures in its learned embedding space as an *emergent* property. As seen, ACE provides mirrored embeddings for mirrored anatomical structures (e.g., the right and left clavicles, and the right and left rib 5.).

B.2. Fine-tuning evaluation: key point detection

Experimental Setup: we investigate the generalizability of ACE's pretrained model via fine-tuning the landmark detection task. To do so, we use the dataset annotated by experts with distinct anatomical landmarks (mentioned in main paper Sec. 5.2), and we choose 7 key points as shown in Fig. 2-a. We load the pretrained weights of ACE and other baselines including ImageNet-1K, BYOL, DINO and POPAR. The fine-tuning architecture is UperNet which is the same with segmentation, while the training target is the specific points of interest. Following [17], we optimize the detection process based on the heatmap method, that is, we add a 11×11 Gaussian kernel $\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$ to smooth each ground truth landmark where the peak is 1 and the values decrease as the distance increase. The learning target is visualized in Fig. 2-b where the green points are the center of the heatmaps. The error between prediction and ground truth points is used as the evaluation metric.

Results: As seen in Fig. 2-c, initializing with our ACE's weights can get the lowest pixel error 16.44 while the image size is 448×448, better than initialized with other baselines, ImageNet pretrained weights and training from scratch. From the results, ACE's representations can give some priors about the anatomical structure which boosts to distinguish the key points.



(a) Chosen key points (b) Key point prediction

Method	Pixel Error ↓
Scratch	29.25
ImageNet-1k	17.36
BYOL	19.79
DINO	19.37
POPAR	17.61
ACE	**16.44**

(c) Comparison on different methods

Figure 2. ACE demonstrates its ability to boost downstream key point detection tasks. (a) 7 key points are chosen for fine-tuning; (b) the inference detection image where the red points are prediction while green points are the ground truth situating at the center of the heatmap; (c) comparison between ACE and other pretrained baselines and the lower pixel error the better detection performance.

B.3. Weakly supervised localization

Experimental setup: To compare with other pretraining methods POPAR [14], DINO [3], BYOL [5] and Adam [8], we initialize downstream model with these pretrained weights using only image-level disease label on ChestX-ray14 dataset. After fine-tuning, the models are used for inference on 787 cases annotated with bounding boxes for eight thorax diseases: Atelectasis, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, and Pneumothorax. We use Grad-CAM [15] heatmap to approximate the localization of a specific thorax disease predicted by the trained model. The baseline Adam is fintuned on ResNet50 and other methods are based on Swin-B.

Results: Fig. 3 shows the visualization of heatmaps generated by ACE, POPAR, DINO, BYOL and Adam for 8 thorax pathologies in ChestX-ray14 dataset. From the results, the localization of our method surpasses the learning global feature methods DINO and BYOL and learning inherent structure pattern method POPAR and Adam. For analyzing the Grad-CAM heatmaps, our method shows more precise and compact localization with small shifts, while the learning global feature methods DINO and BYOL often completely can not localize the diseases. And surprisingly, our model can also localize some small pathologies like nodules and atelectasis, which demonstrate the positive impact of the combination of learning global and local anatomies.

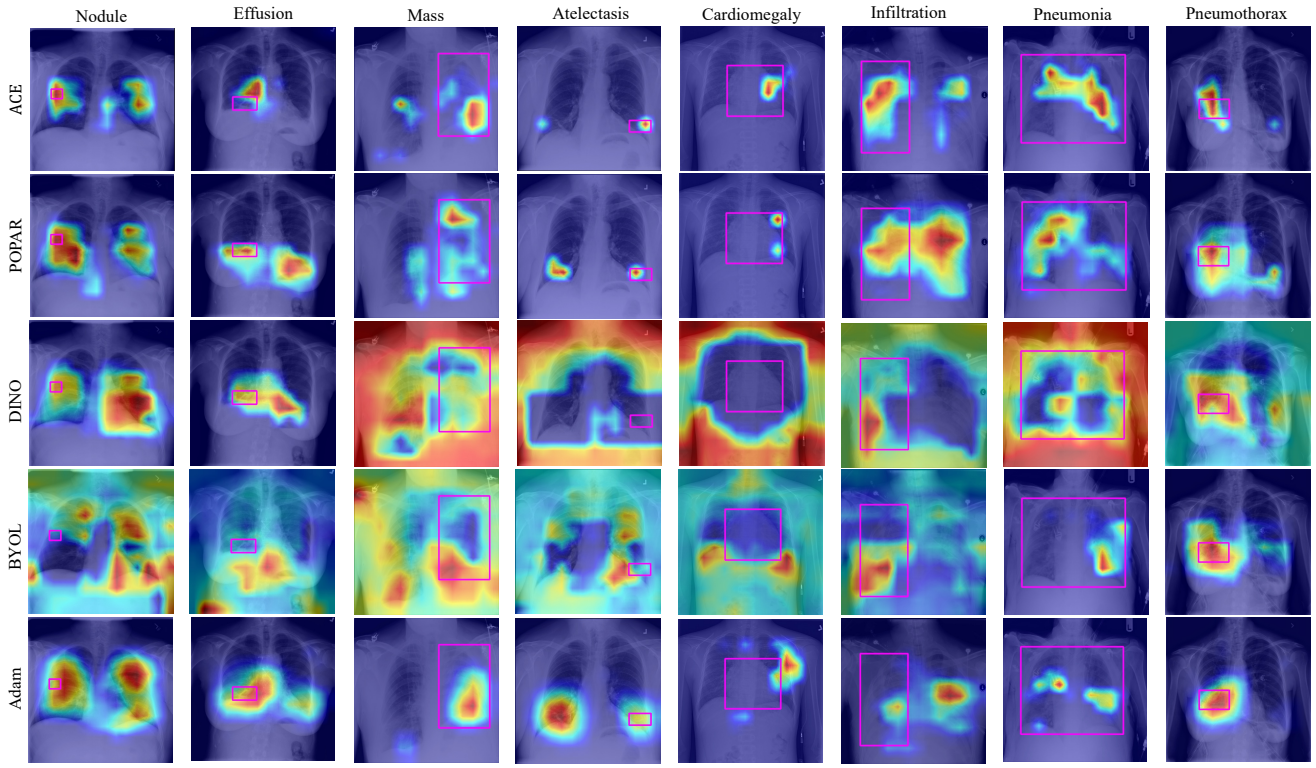


Figure 3. Visualization of Grad-CAM heatmaps. For each column, we provide the heatmap examples for 8 thorax diseases which hold bounding boxes in official labeling. The first row shows the results of our method ACE while the rest rows represent the localization of POPAR, DINO, BYOL and Adam. The pink boxes are the localization ground truth.

References

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

[2] https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation. SIIM-ACR Pneumothorax Segmentation (2019). 2

[3] 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. 3

[4] Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009. 1, 2

[5] 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. 3

[6] Carlos Hernandez-Matas, Xenophon Zabulis, Areti Triantafyllou, Panagiota Anyfanti, Stella Douma, and Antonis A Argyros. Fire: fundus image registration dataset. *Modeling and Artificial Intelligence in Ophthalmology*, 1(4):16–28, 2017. 2

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[8] Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 94–104. Springer, 2023. 3

[9] 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. 1

[10] 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. 1

[11] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*, 2020. 2

[12] 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. 2

[13] 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 *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 12–22. Springer, 2022. 1

[14] 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. 3

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[16] 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. 1

[17] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[18] 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. 1

[19] 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. 2