AFiRe: Anatomy-Driven Self-Supervised Learning for Fine-Grained Representation in Radiographic Images

Yihang Liu [1], Lianghua He [12*], Ying Wen [3], Longzhen Yang [1], Hongzhou Chen [1]

[1] School of Computer Science and Technology, Tongji University, Shanghai, China.
[2] Shanghai Eye Diseases Prevention and Treatment Center, Shanghai Eye Hospital, Shanghai, China.
[3] School of Communication and Electronic Engineering, East China Normal University, Shanghai, China.
{2111131, helianghua, yanglongzhen, chenhonghzou}@tongji.edu.cn, ywen@cs.ecnu.edu.cn

Abstract

Current self-supervised methods, such as contrastive learning, predominantly focus on global discrimination, neglecting the critical fine-grained anatomical details required for accurate radiographic analysis. To address this challenge, we propose the **A**natomy-driven self-supervised framework for enhancing **Fi**ne-grained **Re**presentation in radiographic image analysis (AFiRe). The core idea of AFiRe is to align the anatomical consistency with the unique token-processing characteristics of Vision Transformer. Specifically, AFiRe synergistically performs two self-supervised schemes: (i) Token-wise anatomy-guided contrastive learning, which aligns image tokens based on structural and categorical consistency to enhance fine-grained spatial-anatomical discrimination; (ii) Pixel-level anomaly-removal restoration, which particularly focuses on local anomalies, thereby refining the learned discrimination with detailed geometrical information. Additionally, we propose the Synthetic Lesion Mask to enhance anatomical diversity while preserving intra-consistency, which is typically corrupted by traditional data augmentations, such as Cropping and Affine transformations. Experimental results show that AFiRe: (i) provides robust anatomical discrimination, achieving more cohesive feature clusters compared to state-of-the-art contrastive learning methods; (ii) demonstrates superior generalization, surpassing 7 radiography-specific self-supervised methods in multi-label classification tasks with limited labeling; and (iii) integrates fine-grained information, enabling precise anomaly detection using only image-level annotations.

Code — https://github.com/LYH-hh/AFiRe

Introduction

Contrastive learning (CL) has emerged as a powerful method in self-supervised learning (SSL), demonstrating its effectiveness without relying on expert annotations (Caron et al. 2020, 2021). In natural image analysis, prevailing CL methods like SimCLR (Chen et al. 2020a) and MoCo (He et al. 2020) primarily focus on global discrimination. Explicitly, they consider the entire image and its transformations as positive pairs, while treating different images as negative pairs (Wu et al. 2018; Misra and Maaten 2020). Although

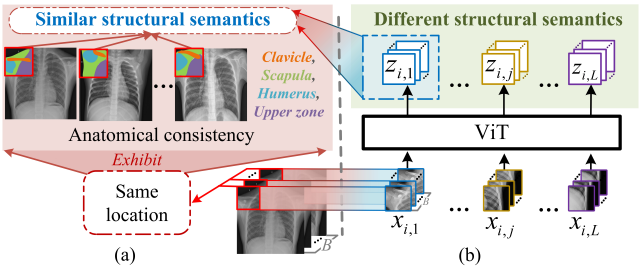


Figure 1: **Conception of the proposed method.** (a) Local radiographic structures at the same location exhibit anatomical consistency. (b) By aligning anatomical consistency with the token-processing characteristics of ViT, tokens at the same position within a batch share similar structural semantics, while those at different positions convey distinct ones.

these models successfully capture high-level representations across diverse images and achieve invariance to salience changes, their efficacy is limited when applied to radiographic images (Zhou et al. 2021a; Haghighi et al. 2022).

Natural images exhibit invariance of foreground objects across diverse backgrounds, which directs contrastive tasks to emphasize image-level discrimination (Misra and Maaten 2020). In contrast, radiographic images present clinical salience dispersed throughout the image (Li et al. 2024), necessitating the model's attention to fine-grained discrimination, including the distribution of densities, the arrangement of tissues, and the presence of specific pathologies (Agu et al. 2021; Zhao et al. 2021b). For example, faint opacities may indicate early signs of pneumonia, while fine reticular patterns can suggest interstitial lung disease. Therefore, incorporating fine-grained details, such as detailed anatomical structures and complex spatial relationships (Haghighi et al. 2024; Chen et al. 2023), into high-level representations is essential for accurately identifying normal anatomy and detecting various lesions in radiographic images.

Recent advances in medical contrastive learning (MCL) have incorporated fine-grained information by focusing on region discrimination (Chen et al. 2023; Singh, Gorade, and Mishra 2024; Li et al. 2024). These approaches aim to embed semantically consistent features while distinguishing them from features with distinct semantics (Taher, Gotway,

*Corresponding author.

and Liang 2024). However, the variability of diseases makes it difficult to comprehensively capture pathological semantics, and additional resources are required to determine the relevant areas (such as pre-selection (Huang et al. 2021)). Another effective method for enriching fine-grained information is to integrate the pixel restoration with MCL (Zhou et al. 2021a; Haghighi et al. 2024). However, these methods typically exhibit a deficiency in anatomical semantics, as they restore pixel-level content directly from latent representations. This plain pixel information limits the model's capability to focus on salient clinical patterns.

In this paper, to circumvent the challenge posed by heterogeneous anomalies, which potentially leads to insufficient learning of semantic information, we highlight that *comprehensive learning of normal anatomical patterns is comparatively easier.* This advantage stems from the significant semantic similarity observed in normal radiographic images (Xiang et al. 2023), which are characterized by anatomical consistency and stable focal regions across different individuals (Fig. 1(a)). Such similarity facilitates the effective decoupling of underlying anomalies from normal structures, thereby enhancing pathological diagnosis performance. Additionally, to avoid region pre-selection, we hypothesize that *each distinct token in Vision Transformers (ViTs) (Dosovitskiy et al. 2021) explicitly represents the predominant information of specific local anatomical structures*. This hypothesis is inspired by SelfPatch (Yun et al. 2022), which asserts that ViTs have inherent architectural advantages for enhancing visual representations through the processing of discrete image tokens. Summarily, our core idea is to ***align anatomical consistency with the unique token-processing characteristics of ViT*** to enhance fine-grained radiographic representation (Fig. 1(b)). These alignments are also pivotal in anomaly-specific restoration for preserving pixel-level anatomy-associated (*i.e.*, geometrical) information.

Based on the analysis above, we propose a novel **A**natomy-driven self-supervised framework to enhance **Fi**ne-grained **Re**presentation in radiographic images (**AFiRe**). Equipped with the designed Synthetic Lesion Masks (SLMs), an anatomy augmentation inspired by AnatPaste (Sato et al. 2023), AFiRe aligns ViT tokens with local anatomical structures via synergistically performing two SSL tasks: Token-wise anatomy-guided contrastive learning and Pixel-level anomaly-removal restoration.

For the contrastive learning task, instead of analyzing entire images, AFiRe processes each disjoint token in ViT as an independent sample. By maximizing mutual information among these tokens within a batch, this task learns fine-grained structural invariance and discriminative semantics based on structural and categorical consistency. To guide the model in comprehensively learning the normal anatomical patterns, we introduce a group of spatial-aware prototypes in this task. These prototypes represent the distribution of different anatomical structures, serving as pseudo-cluster assignments for the predicted token probabilities. For the pixel restoration task, we strategically focus on restoring specific abnormal tokens augmented by SLMs. Concretely, we remove these abnormal tokens by substituting them with trainable masked tokens to restore their corresponding normal pixels. Hence, the latent representations retain both normal and abnormal information while preserving geometrical details closely associated with various anatomical structures.

Our extensive experiments demonstrate that AFiRe enhances the model's capacity to capture fine-grained discriminative information from normal radiographic images, facilitating a more robust representation across different anatomic structures. Compared to different supervised and self-supervised benchmarks, AFiRe achieves superior performance in multi-label classification tasks, indicating its generalization in analyzing real disease images. Furthermore, AFiRe has observable advantages in three anomaly detection tasks, outperforming the state-of-the-art counterparts. Overall, our contributions are as follows:

- We design a Token-wise anatomy-guided contrastive learning SSL task, equipped with spatial-aware prototypes to integrate fine-grained anatomical discrimination based on structural and categorical consistency.
- We design a Pixel-level anomaly-removal restoration SSL task to preserve detailed geometrical information closely associated with various anatomical structures.
- We introduce Synthetic Lesion Mask, an efficient anatomical data augmentation technique, to enhance anatomical diversity while preserving intra-consistency.
- We propose an Anatomy-driven self-supervised framework that synergistically optimizes the aforementioned tasks to achieve robust radiographic representation with fine-grained anatomical information.

Related Work

SSL has become a prominent topic in medical image analysis, enabling the extraction of meaningful representations directly from data without the need for explicit disease-specific labels. In this section, we review recent advancements that are most relevant to our proposed AFiRe method.

Medical Contrastive Learning. This approach typically uses encoders to cluster instances within the pseudo-classes (Sowrirajan et al. 2021; Azizi et al. 2021; Vu et al. 2021). Recent research, including C2L (Zhou et al. 2020) adapts general contrastive methods to learn distinctive patterns across various medical images. Adamv1 (Hosseinzadeh Taher, Gotway, and Liang 2023) applies hierarchical contrastive learning, capturing anatomy in a coarse-to-fine manner. Adamv2 (Taher, Gotway, and Liang 2024) expands it by enhancing radiographic representation with localizability, composability, and decomposability. While these methods show effective applications in medical tasks, they remain focused on image-level discrimination based on the input, neglecting the fine-grained anatomical information that could be aligned with token-level representations.

Medical Restorative Learning. This approach reconstructs original images from corrupted versions, enhancing pixel-level details to identify normal anatomy and various abnormalities (Haghighi et al. 2024). Key advancements include Model Genesis (MG) (Zhou et al. 2021b), which uses image restoration tasks on unlabeled medical images; TransVW (Haghighi et al. 2021), which defines a "visual word" as a

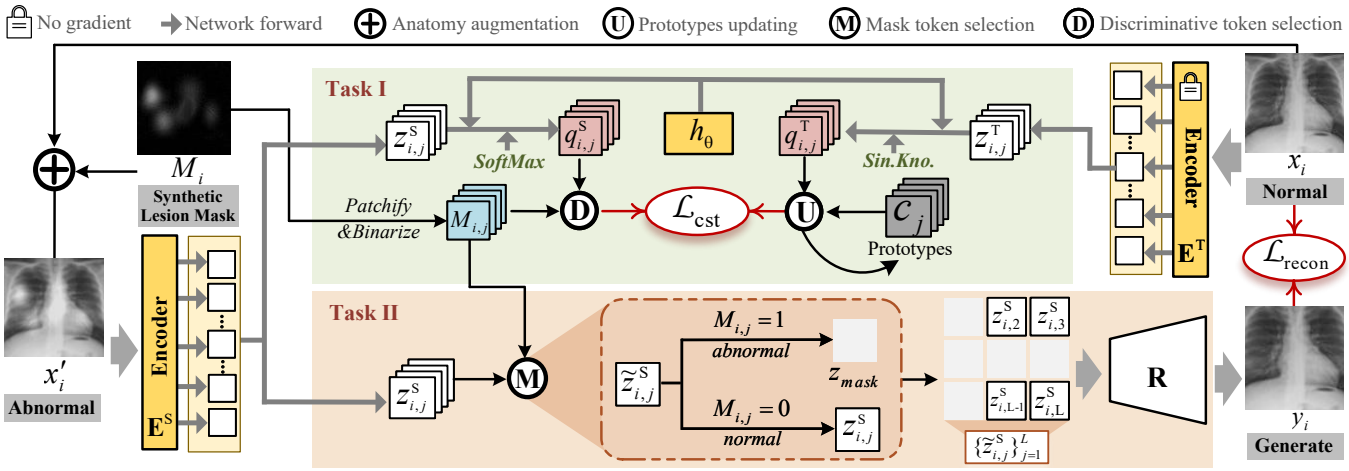


Figure 2: **Overview of the proposed AFiRe**. It synergistically performs two self-supervised proxy tasks: Token-wise anatomy-guided contrastive learning (**Task I**) and Pixel-level anomaly-removal restoration (**Task II**). For each normal input x_i, we perturb it using the designed Synthetic Lesion Mask (M_i) to produce abnormal input x_i'. In Task I, a group of spatial-aware prototypes, updated by the teacher network's output, serve as pseudo-cluster labels to maximize alignment among tokens from student networks belonging to the same class or structure. In Task II, the restoration target particularly focuses on the abnormal tokens from augmented pairs of normal radiographic images by substituting them with mask tokens in the latent space.

recurring anatomical segment across images. SQUID (Xiang et al. 2023) leverages space-aware memory queues to capture spatial correlations and consistent anatomical structures in chest images, thereby demonstrating effectiveness in anomaly detection through unsupervised learning. AnatPaste (Sato et al. 2023) introduces anatomy-aware pasting augmentation, generating synthetic images to distinguish real normal images with a one-class classifier under SSL. These methods have shown promise in radiographic representation learning by preserving detailed patterns; however, they are often suboptimal for radiography classification due to limited high-level discrimination.

Medical Synergistic Learning. To harness complementary advantages, synergistic SSL approaches have been proposed for radiographic representation learning. DiRA (Haghighi et al. 2022) combines discriminative, restorative, and adversarial learning to capture complementary visual information, enhancing fine-grained semantic representation learning. PCRL (Zhou et al. 2021a) incorporates preservation mechanisms to reconstruct diverse image contexts, refining the representations derived from contrastive methods.

While these models offer progress, the proposed AFiRe advances the comprehensive integration of anatomical information by introducing Token-Wise Anatomy-Guided Contrastive Learning and Pixel-Level Anomaly-Removal Restoration, achieving both fine-grained high-level discrimination and preservation of pixel-level anatomical geometry.

Method

To enhance fine-grained anatomical representation in radiographic image analysis, we propose an anatomy-driven self-supervised framework as depicted in Fig. 2. Our framework employs a siamese ViT architecture to extract features from real normal and generated abnormal images by introducing the Synthetic Lesion Mask. Based on the alignment of anatomical consistency with the unique token-processing characteristics of ViT, this framework synergistically performs two SSL tasks: (i) **Task I**, the Token-wise anatomy-guided contrastive learning, aims to enhance fine-grained discriminative information by maximizing mutual information among individual image tokens within the same class or anatomical structure; and (ii) **Task II**, the Pixel-level anomaly-removal restoration, aims to incorporate geometric-based anatomical details by optimizing the alignment of abnormal tokens between the normal image and its reconstructed counterpart at the pixel level.

In the following, we introduce each component individually and then discuss the synergistic training scheme.

Feature Extraction with Siamese Network

As shown in Fig. 2, the overall structure of the AFiRe comprises a student branch and a teacher branch. In our training scheme, the student network $\mathbf{E}^{\mathrm{S}}$, parameterized by θ^{S}, is trained to match the spatial-aware prototypes updated by the output of the teacher network $\mathbf{E}^{\mathrm{T}}$, parameterized by θ^{T}. Both networks share identical ViT architectures. During the pre-training phase, only θ^{S} is updated via back-propagation, while θ^{T} is updated using an exponential moving average (EMA) (He et al. 2020; Caron et al. 2021) of θ^{S} as follows:

$$\theta^{\mathrm{T}} \leftarrow \lambda\theta^{\mathrm{T}} + (1-\lambda)\theta^{\mathrm{S}}, \tag{1}$$

where λ follows a cosine schedule from 0.99 to 1.

To leverage the anatomical consistency observed in normal radiographic images (Xiang et al. 2023), we incorporate two types of input data during the pre-training stage: (i) Real normal images $\{x_i\}_{i=1}^{B}$, which provide a baseline to establish the distribution of typical anatomical structures, and (ii) Synthetic abnormal images $\{x_i'\}_{i=1}^{B}$, which integrate category discriminative information by simulating pathologies

using Synthetic Lesion Masks. Here, B represents the batch size, and the images x_i and x_i' are encoded by the teacher network $\mathbf{E}^{\mathrm{T}}$ and the student network $\mathbf{E}^{\mathrm{S}}$, respectively.

Alignment Anatomical Consistency with ViT Tokens. ViTs process input images as sequences of non-overlapping patches, with the extracted features $\{z_{i,j}^{\mathrm{T}}\}_{j=1}^{L}$ and $\{z_{i,j}^{\mathrm{S}}\}_{j=1}^{L}$ (white blocks in Fig. 2) corresponding to token-wise representations, where L is the length of image token sequences. Each $z_{i,j}$ represents the j-th positional token, corresponding to the j-th local anatomical structure in the i-th image. In the subsequent Task I and Task II, we explicitly align the anatomical consistency exhibited in $\{z_{i,j}^{\mathrm{T}}\}_{i=1}^{B}$ and $\{z_{i,j}^{\mathrm{S}}\}_{i=1}^{B}$ with the unique token-processing characteristics of ViTs to enhance fine-grained radiographic representation.

Anatomy Augmentation via Synthetic Lesion Mask. Traditional image data augmentations, such as geometric transformations including Rotation, Cropping, Affine, and Flipping, tend to corrupt anatomical consistency due to altered spatial relationships. Popular pasting-based augmentations like CutPaste (Li et al. 2021) are designed to introduce variations in images without causing geometric distortions. However, these techniques often produce abnormalities with clear boundaries, which fail to capture the effective characteristics of real anomalies. Recently, the anatomy-aware pasting (AnatPaste) augmentation technique (Sato et al. 2023) has been developed to create anomalies with anatomical fidelity by introducing abnormal shadows within the extracted lung regions. In this section, we improve AnatPaste in an efficient way to generate more random anomalies by introducing the Synthetic Lesion Mask (SLM).

Unlike AnatPaste, which uses fixed anomalous sizes and textures, SLM implements a dynamic anatomical augmentation to efficiently synthesize various pathological anomalies within a single normal image. Specifically, SLM randomly selects a highlighted region η, binarized by (Otsu et al. 1975), from x_i to serve as universal texture noise. For the size of $x_i \in \mathbb{R}^{224\times224}$, we simulate lesion sizes ranging from small nodules to large consolidations by randomly resizing η to $\mathbb{R}^{W\times H}$, where $W, H \in [16, 64]$. We then define the lesion shapes as irregular ovals:

$$\delta(w,h) = \exp(-\frac{\rho(w,h)}{\gamma}^2), \tag{2}$$

$$\rho(w,h) = \sqrt{(w-\frac{W}{2})^2 + (h-\frac{H}{2})^2}. \tag{3}$$

In the shape $\delta(w,h) \in \mathbb{R}^{W\times H}$, the value at each pixel (w,h) gradually decreases from the center to the edge with the radius γ. We integrate η with $\delta(w,h)$ and randomly put them onto a zero-valued map of the same size as x_i. The process of creating a group of SLMs $\{M_i^n\}_{n=1}^{N}$ for each x_i is formulated as follows:

$$\{M_i^n\}_{n=1}^{N} = \{\sum_{r=1}^{R} \eta^r \cdot \delta^n(w,h)\}_{n=1}^{N}, \tag{4}$$

where N is the number of SLMs, and R represents the number of abnormal regions in a single M_i^n. Empirically, we set $R \in [1, 4]$. The abnormal images can be obtained by:

$$\{x_i'^{,n}\}_{n=1}^{N} = \{x_i \oplus M_i^n\}_{n=1}^{N}, \tag{5}$$

where $\oplus$ is the element-wise additional operation to augment the anatomy.

Token-wise Anatomy-guided Contrastive Learning

Instead of contrasting the image-level representation, in this task, we consider each token as an individual sample. Specifically, we propose Token-wise anatomy-guided contrastive learning (**Task I**), which maximizes the mutual information between the probability distribution of normal image tokens and their respective structural prototypes (Fig. 3) by integrating structure-consistency and category-consistency contrastive losses (Fig. 4).

Token Probabilities Prediction. To mitigate the impact of individual anatomical differences, our contrastive learning task focuses on cluster assignments rather than direct feature comparisons. Specifically, we employ a projection head (h_θ) to map each token from the dim-dimensional feature ($\{z_{i,j}^{\mathrm{S}}\}$ and $\{z_{i,j}^{\mathrm{T}}\}$) to K-dimensional vectors as shown in Task I of Fig. 2. The student token probability distributions $q_{i,j}^{\mathrm{S}} \in \mathbb{R}^K$ are obtained by normalizing these K-dimensional vectors with a $SoftMax$ function, while the teacher token probability distributions $q_{i,j}^{\mathrm{T}} \in \mathbb{R}^K$ are computed via the Sinkhorn-Knopp ($Sin.Kno.$) algorithm (Cuturi 2013) to avoid trivial solutions where all samples collapse into a single representation. Here, the $q_{i,j,k}$, where $k \in [1, K]$, denotes the predicted probability that the j-th token of the i-th image belongs to class k within its structural cluster assignment.

Discriminative Token Selection. To achieve consistency among normal anatomic structures, it is important to identify probability distributions in normal and abnormal categories. Synthetic Lesion Mask M_i is regarded as the token label to distinguish normal and abnormal image tokens within the set $\{q_{i,j}^{\mathrm{S}}\}_{j=1}^{L}$, corresponding to the '**D**' operation in Fig. 2. Specifically, M_i is patchified with the same patch size as x_i, resulting in a sequence $\{M_{i,j}\}_{j=1}^{L}$. We define the threshold as the average pixel value of M_i. Therefore, for each patch, $M_{i,j} = 1$ indicates that the patch's average pixel value exceeds the threshold; otherwise, $M_{i,j} = 0$. Thus, the image tokens can be collected in two categories:

$$\{q_{i,j}^{\mathrm{S},+}\} = \{q_{i,j}^{\mathrm{S}} | M_{i,j} = 0\},\ \{q_{i,j}^{\mathrm{S},-}\} = \{q_{i,j}^{\mathrm{S}} | M_{i,j} = 1\}, \tag{6}$$

where $q_{i,j}^{\mathrm{S},+}$ and $q_{i,j}^{\mathrm{S},-}$ indicate the normal and abnormal token probability, respectively.

Spatial-aware Prototypes Construction and Update. To learn comprehensive semantics of various normal anatomical structures, we adopt the concept of clustering prototypes from SwAV (Caron et al. 2020). Our approach diverges significantly from SwAV by introducing a group of distinct prototype vectors (the gray blocks in Fig. 2), allowing the model to more effectively represent the probabilistic distribution across different anatomies. In our pre-training stage, these prototypes act as the pseudo cluster assignments for $q_{i,j}^{\mathrm{S}}$, thereby guiding the update of gradient for θ^{S}.

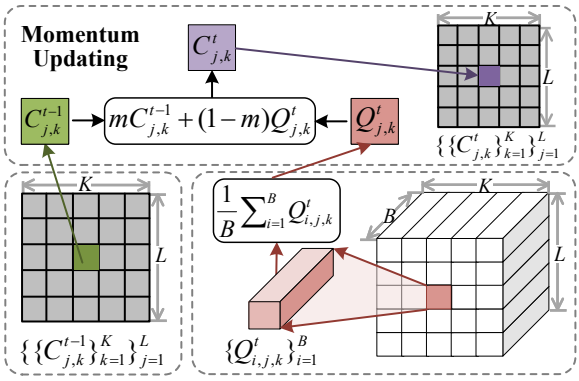


Figure 3: **Updating process of the spatial-aware prototypes**. The cluster assignment of $\mathbf{E}^{\mathrm{T}}$ is used for updating the spatial-aware prototypes.

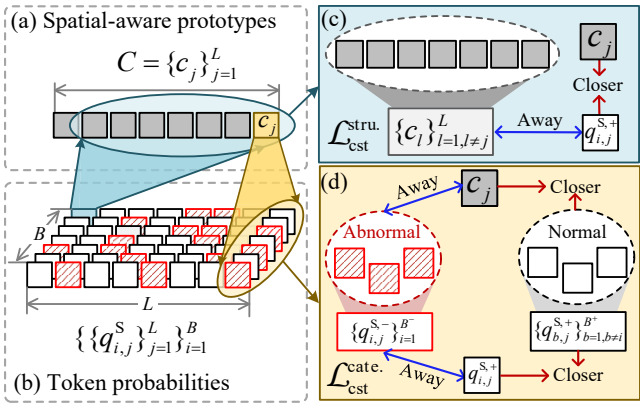


Figure 4: **Token-wise anatomy-guided contrastive learning**. $\mathcal{L}^{\mathrm{stru.}}_{\mathrm{cst}}$ and $\mathcal{L}^{\mathrm{cate.}}_{\mathrm{cst}}$ correspond to the structure-consistency and category-consistency contrastive losses, respectively.

In practice, these vectors are learned through momentum updating rather than back-propagation (the '**U**' operation in the Fig. 2). We define a 2D prototype matrix $\mathbf{C} = \{\{c_{j,k}\}^K_{k=1}\}^L_{j=1} \in \mathbb{R}^{L\times K}$ as shown in Fig. 3. This matrix indicates that each distinct token in the image is described by a K-dimensional vector. At time t, the average probability score for the specific j-th token $\{q^{t,\mathrm{T}}_{i,j,k}\}^B_{i=1}$ in the k-th class is calculated as $q^t_{j,k} = \frac{1}{B}\sum_{i=1}^{B}(q^{t,\mathrm{T}}_{i,j,k})$ (the pink block in Fig. 3). The updating of previous prototypes $c^{t-1}_{i,k}$ (the green block in Fig. 3) can be formulated as:

$$\begin{cases} c^t_{j,k} = q^t_{j,k}, & \text{if } t=1 \\ c^t_{j,k} = mc^{t-1}_{j,k} + (1-m)q^t_{j,k}, & \text{if } t>1. \end{cases} \tag{7}$$

This process ensures that the prototype matrix $\mathbf{C}$ is gradually updated with the predicted probabilities over time, balancing between the historical values and the new incoming data through the momentum parameter m.

Structure-consistency Contrastive Loss Formulation. To encode more fine-grained structural discriminative information, this contrastive loss is to maintain invariance within the same anatomical structure and increase differentiation among various anatomical structures. Specifically, given the normal token probabilities $Q^{\mathrm{S},+}_i = \{q^{\mathrm{S},+}_{i,j}\}^L_{j=1}$ in the i-th image (such as the blue block in Fig. 4 (b)) and the spatial-aware prototypes $C = \{c_j\}^L_{j=1}$, we maximize the mutual information as follows:

$$I(Q^{\mathrm{S},+}_i; C) = \sum_{q^{\mathrm{S},+}_{i,j}}\sum_{c_j} P(c_j, q^{\mathrm{S},+}_{i,j}) \log \frac{P(c_j|q^{\mathrm{S},+}_{i,j})}{P(c_j)}. \tag{8}$$

Directly optimizing $I(Q^{\mathrm{S},+}_i; C)$ is a challenging task. Alternatively, we define the joint probability by considering the likelihood that a token $q^{\mathrm{S},+}_{i,j}$ belongs to a cluster represented by c_j. The conditional probability is approximated by the normalized similarity between them. This process can be formulated as follows:

$$P(c_j, q^{\mathrm{S},+}_{i,j}) = P(c_j|q^{\mathrm{S},+}_{i,j})P(q^{\mathrm{S},+}_{i,j}), \tag{9}$$

$$P(c_j|q^{\mathrm{S},+}_{i,j}) = \frac{\mathrm{sim}(q^{\mathrm{S},+}_{i,j}, c_j)}{\sum_{l=1}^{L}\mathrm{sim}(q^{\mathrm{S},+}_{i,j}, c_l)}, \tag{10}$$

where $\mathrm{sim}(.,.)$ is the similarity function. Next, we plug Eq. 9 and Eq. 10 into Eq. 8 and simplify it to obtain:

$$I(Q^{\mathrm{S},+}_i; C) = \mathbb{E}_{P(Q^{\mathrm{S},+}_i, C)}\left[\log \frac{\mathrm{sim}(q^{\mathrm{S},+}_{i,j}, c_j)}{\sum_{l=1}^{L}\mathrm{sim}(q^{\mathrm{S},+}_{i,j}, c_l)}\right]. \tag{11}$$

Therefore, $I(q^{\mathrm{S},+}_{i,j}; c_j) \propto \mathrm{sim}(q^{\mathrm{S},+}_{i,j}, c_j)$. By considering all the $i \in [1, B^+]$, and $j \in [1, L]$, we maximize $I(Q^{\mathrm{S},+}_i; C)$ by proposing the structure-consistency contrastive loss:

$$\mathcal{L}^{\mathrm{stru.}}_{\mathrm{cst}} = -\frac{1}{LB^+}\sum_{j=1}^{L}\sum_{i=1}^{B^+}\log\frac{f(q^{\mathrm{S},+}_{i,j}, c_j)}{\sum_{l=1}^{L} f(q^{\mathrm{S},+}_{i,j}, c_l)}, \tag{12}$$

where $f(q,c) = \exp(\mathrm{H}(q,c)/\tau)$ and B^+ is the number of $q^{\mathrm{S},+}_{i,j}$. Here, τ is the temperature parameter and $\mathrm{H}(q,c) = -\sum_{k=1}^{K} q(k)\log c(k)$, which is the cross-entropy function to measure the similarity between two probabilities. The positive and negative pairs in $\mathcal{L}^{\mathrm{stru.}}_{\mathrm{cst}}$ are defined as depicted in Fig. 4 (c). We pull a specific normal image token $q^{\mathrm{S},+}_{i,j}$ and its corresponding structural prototype c_j closer while pushing other prototypes $\{c_l\}^L_{l=1}$, where $l \neq j$, away.

Category-consistency Contrastive Loss Formulation. To effectively decouple pathological information from the normal distribution, we further investigate the consistency among normal image tokens within a batch size B at the same position j. Particularly, given the specific $c_j \in C$ (the yellow block in Fig. 4 (a)) and $q^{\mathrm{S},+}_{i,j}, q^{\mathrm{S},+}_{b,j} \in Q^{\mathrm{S},+}_j$, we maximize the joint mutual information as follows:

$$I(C, Q^{\mathrm{S},+}_{j^b}; Q^{\mathrm{S},+}_{j^i}) = I(C; Q^{\mathrm{S},+}_{j^i}) + I(Q^{\mathrm{S},+}_{j^i}; Q^{\mathrm{S},+}_{j^b}|C), \tag{13}$$

where $i, b \in [1, B^+]$ and $b \neq i$. Since $I(C; Q_{j^i}^{\mathrm{S},+})$ has been optimized through $\mathcal{L}_{\text{cst}}^{\text{stru.}}$, maximizing Eq. 13 only needs to optimize the second term, which represents the unique mutual information between $q_{b,j}^{\mathrm{S},+}$ and $q_{i,j}^{\mathrm{S},+}$. Thereby, we introduce a negative term, *i.e.*, reducing the similarity between $q_{i,j}^{\mathrm{S},-}$ and c_j, to constrain $I(Q_{j^i}^{\mathrm{S},+}; Q_{j^b}^{\mathrm{S},+}|C)$. Mathematically, similar to Eq. 12, we propose the category-consistency contrastive loss $\mathcal{L}_{\text{cst}}^{\text{cate.}}$ and introduce positive and negative terms as depicted in Fig. 4 (d):

$$\Delta(pos) = \sum\nolimits_{b=1}^{B^+} f(q_{i,j}^{\mathrm{S},+}, q_{b,j}^{\mathrm{S},+}), \tag{14}$$

$$\Delta(neg) = \sum\nolimits_{b=1}^{B^-} f(q_{i,j}^{\mathrm{S},+}, q_{b,j}^{\mathrm{S},-}) + f(q_{i,j}^{\mathrm{S},-}, c_j), \tag{15}$$

where B^- is the number of $q_{i,j}^{\mathrm{S}}$ and $B^+ + B^- = B$. Therefore, the $\mathcal{L}_{\text{cst}}^{\text{cate.}}$ can be formulated as follows:

$$\mathcal{L}_{\text{cst}}^{\text{cate.}} = -\frac{1}{LB} \sum_{j=1}^{L} \sum_{i=1}^{B} \log \frac{\Delta(pos)}{\Delta(pos) + \Delta(neg)}. \tag{16}$$

Notably, given the inherent diversity and irregularity of pathologies in radiographic images, enforcing consistency among abnormal tokens could limit the model's ability to generalize and recognize pathological semantic information. Therefore, in the proposed $\mathcal{L}_{\text{cst}}^{\text{cate.}}$, we disregard intra-class similarity for the $I(Q_{jb}^{\mathrm{S},-}, Q_{ji}^{\mathrm{S},-})$, where $q_{i,j}^{\mathrm{S},-}, q_{b,j}^{\mathrm{S},-} \in Q_j^{\mathrm{S},-}$, to accommodate the heterogeneity of anomalies.

Summarily, in the proposed Token-wise anatomy-guided contrastive task, the loss function is formulated as follows:

$$\mathcal{L}_{\text{cst}} = \mathcal{L}_{\text{cst}}^{\text{stru.}} + \mathcal{L}_{\text{cst}}^{\text{cate.}}. \tag{17}$$

Pixel-level Anomaly-removal Restoration

To enhance the learned discrimination with detailed geometrical information, we propose the pixel-level anomaly-removal restoration, as illustrated in the **Task II** of Fig. 2. Unlike conventional methods that apply image restoration across the entire latent feature space, this task concentrates on localized anomalies by restoring image features that exclude abnormal tokens. Specifically, $\{M_{i,j}\}_{j=1}^{L}$ is utilized to identify anomalies within the latent token features $z_{i,j}^{\mathrm{S}}$ (*i.e.*, '**M**' operation in the Fig. 2). We introduce trainable mask tokens, z_{mask}, to replace these abnormal tokens in the latent feature space. Consequently, given the decoder $\mathbf{R}$, the restored image can be obtained as follows:

$$\{\{y_{i,j}\}_{j=1}^{L}\}_{i=1}^{B} = \mathbf{R}(\{\{\tilde{z}_{i,j}^{\mathrm{S}}\}_{j=1}^{L}\}_{i=1}^{B}), \tag{18}$$

$$\tilde{z}_{i,j}^{\mathrm{S}} = \begin{cases} z_{i,j}^{\mathrm{S}}, & \text{if } M_{i,j} = 0 \\ z_{mask}, & \text{if } M_{i,j} = 1. \end{cases} \tag{19}$$

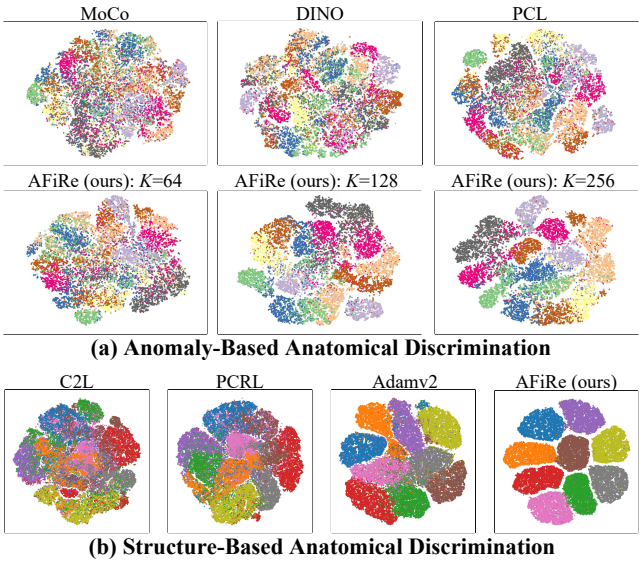
To improve the quality of generated abnormal regions while preserving anatomical details, we introduce a weighted mean squared error (MSE) loss function:

$$\mathcal{L}_{\text{recon}} = \frac{1}{L} \sum_{j=1}^{L} w_j \|y_{i,j} - x_{i,j}\|^2, \tag{20}$$

where w_j is the weighted value to emphasize anomalies.

#	Aug.	*Proto.*	$\mathcal{L}_{\text{cst}}^{\text{cate.}}$	$\mathcal{L}_{\text{cst}}^{\text{stru.}}$	$\mathcal{L}_{\text{recon}}$	AUC	ACC	F1
1	SLM	✔	✔	✔	✔	**89.4**	**83.8**	**84.5**
2	AnatPaste	✔	✔	✔	✔	88.9	83.2	83.6
3	SLM	✗	✔	✔	✔	83.1	75.8	79.7
4	SLM	✔	✗	✗	✔	76.9	67.2	75.9
5	SLM	✔	✗	✔	✔	87.7	79.8	78.4
6	SLM	✔	✔	✗	✔	84.5	74.9	76.9
7	SLM	✔	✔	✔	✗	86.2	80.8	81.5

Table 1: **Impact of different components.** AUC scores ↑ (%), ACC scores ↑ (%), F1 scores ↑ (%) on *Child*CXR.



(a) Anomaly-Based Anatomical Discrimination

(b) Structure-Based Anatomical Discrimination

Figure 5: **T-SNE visualizations of the learned representation.** Different colors represent various anomalies (disease classes) in (a) and different image locations in (b).

Overall Training Objective

Our training methodology synergistically optimizes the aforementioned losses to enhance fine-grained anatomical discrimination in radiographic representation. The overall training objective is delineated as follows:

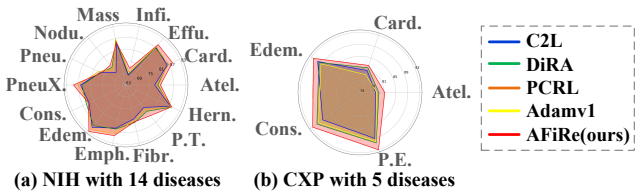
$$\mathcal{L}(x_{i,j}, x'_{i,j}) = \mathcal{L}_{\text{cst}}(x_{i,j}, x'_{i,j}) + \mathcal{L}_{\text{recon}}(x_{i,j}, x'_{i,j}). \tag{21}$$

Experiment

Implementation Details

Pre-training Dataset. In this paper, we assemble a dataset of 811,170 Chest X-ray (CXR) images for pre-training, which includes 81,117 normal CXR images sourced from MIMIC-CXR-JPG (Johnson et al. 2019) and $81,117 \times 9$ synthetic abnormal images. Each normal CXR image is employed to generate nine synthetic abnormal images.

Fine-tuning Datasets. We evaluate our model on three CXR datasets: (i) ***Child*CXRs** (Kermany et al. 2018), (ii) **NIH** (Wang et al. 2017), (iii) **CXP** (Irvin et al. 2019), and



(a) NIH with 14 diseases (b) CXP with 5 diseases

Figure 6: **Comparison of AUC scores ↑ (%) for each disease** on the NIH dataset and CXP dataset using five-fold cross-validation.

(iv) **SIIM-ARC** [1], using the official data partitions.

Training Settings. We follow the DINO (Caron et al. 2021) training paradigm (*e.g.*, optimizer, learning rate, and weight decay schedule), utilizing an input image size of 224×224. For model evaluation, we use the *pre-training (on source data)* $\rightarrow$ *fine-tuning (on target data)* protocol and employ two settings: transfer learning and anomaly detection. We implement our pre-training model using PyTorch and distribute the training across 8 NVIDIA A6000 GPUs, running for 800 epochs with a batch size of 64 for 6 days.

Ablation Study

Table 1 elucidates the impact of various components in the proposed AFiRe. We evaluate the linear probing performance on the *Child*CXR dataset by reporting AUC, ACC, and F1 scores. The components analyzed include different augmentation techniques (SLM vs. AnatPaste), the integration of the spacial-aware prototypes (*Proto.*), and different pre-training proxy tasks such as category-consistency contrastive learning ($\mathcal{L}_{\text{cst}}^{\text{cate.}}$), structure-consistency contrastive learning ($\mathcal{L}_{\text{cst}}^{\text{stru.}}$), and anomaly-removal pixel restoration ($\mathcal{L}_{\text{recon}}$). The results reveal that our proposed model achieves peak performance with an AUC of 89.4%, ACC of 83.8%, and F1 of 84.5% when all components are active (row 1).

Substituting SLM with AnatPaste (row 2) leads to a slight performance drop across all metrics, indicating SLM's superior benefits and simplicity in generating pseudo lesions due to random η and $\delta(w, h)$ selection. The suboptimal results in row 3 underline the necessity of spatial-aware prototypes for fine-grained anatomical discrimination. Rows 4, 5, and 6 highlight the importance of the proposed contrastive loss, as the removal of any one of them results in a significant performance decrease. Row 7 shows a notable performance decrease when the restoration task is omitted, underscoring the importance of pixel-level information for radiographic image analysis.

Comparisons with State of the Arts

AFiRe achieves robust anatomical discrimination. We use t-SNE (Van der Maaten and Hinton 2008) to visualize the learned representations of AFiRe and other comparable methods, focusing on (i) anomaly-based and (ii) structure-based anatomical discrimination. For experiment (i), we leverage pre-trained parameters from MoCo (Chen et al. 2020b), DINO (Caron et al. 2021), PCL (Li et al. 2020), and our AFiRe under varying cluster numbers (K) to extract radiographic representations. As shown in Fig. 5 (a), AFiRe achieves superior clustering performance for anomalies across 12 disease classes compared to other methods, attributed to its token-wise representation contrast enhanced by spatial-aware prototypes. The distinct separation of clusters becomes increasingly evident with higher values of K, *indicating the effectiveness of AFiRe in discriminating complex anatomical features in real-world diseases.* This observation aligns with DINO's finding that a larger output dimensionality enhances performance. For experiment (ii), we adopt the settings of Adamv2 to annotate different image localizations (9 patches) as distinct anatomical landmarks. AFiRe is compared with three SOTA medical SSL methods: C2L (Zhou et al. 2020), PCRL (Zhou et al. 2021a), and Adamv2 (Taher, Gotway, and Liang 2024). As shown in Fig. 5(b), AFiRe achieves more cohesive feature clustering in structural anatomical discrimination. This result highlights AFiRe's ability to *effectively differentiate various fine-grained anatomical structures.*

AFiRe demonstrates effective generalization. We evaluate the generalization ability of AFiRe by comparing it with 11 pre-training methods, reporting AUC scores for multi-label classification and Dice scores for segmentation across different labeling ratios (1%, 10%, and 100%) on three downstream datasets. To ensure robustness, we further conduct five-fold cross-validation for these experiments.

The results of the comparison are presented in Table 2. AFiRe was validated on NIH, CXP, and SIIM-ACR against: (i) a supervised baseline (ViTB pre-trained on ImageNet), demonstrating a significant improvement in radiographic image analysis ($+4.6\%$ AUC on NIH, $+3.7\%$ AUC on CXP and $+19.2\%$ Dice on SIIM-ACR) when using our SSL tasks; (ii) various SSL methods pre-trained on ImageNet, including MoCoV2, DINO, and MAE, with AFiRe achieving statistically significant improvements ($p < 0.01$); and (iii) SSL methods specifically pre-trained on radiographic images, e.g., C2L, TransVW, MG, DiRA, PCRL, Adamv1, and Adamv2. AFiRe consistently outperforms these radiography-specific methods, especially at lower labeling ratios, highlighting its generalization in data-scarce scenarios. Moreover, Fig. 6 provides a detailed comparison of the AUC scores for each disease across the NIH and CXP datasets. In medical imaging, Cardiomegaly typically maintains well-defined borders despite an enlarged silhouette, while Edema and Emphysema present with varying degrees of increased or decreased density and often exhibit blurred or ill-defined borders. Our model excels in detecting these diseases, outperforming recent state-of-the-art methods with top AUC scores of 89.4%, 93.4%, and 89.6% on NIH, respectively. *These results illustrate that AFiRe can be generalized to real disease types by pre-training with the SLM-augmented synthetic images.*

For the models in (i) and (ii), we re-implement them on NIH, CXP, and SIIM-ACR using publicly available ViT-B pre-trained parameters and the same data splits. For meth-

[1] https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

Method	NIH				CXP				SIIM-ACR			
	1%	10%	100%	5-C	1%	10%	100%	5-C	1%	10%	100%	5-C
ViTB (Dosovitskiy et al. 2021)	57.1	67.5	80.1	79.9	75.1	84.0	87.4	85.3	34.9	56.9	73.2	72.3
MoCov2 (Chen et al. 2020b)	59.2	68.9	80.8	80.6	76.2	84.3	87.0	86.3	30.7	53.7	74.8	72.5
DINO (Caron et al. 2021)	60.6	69.6	81.0	80.3	76.9	84.9	87.6	86.1	35.8	57.4	77.4	74.8
MAE (He et al. 2022)	60.5	70.1	80.9	81.7	77.7	85.2	88.0	86.5	41.2	60.9	79.3	79.1
C2L (Zhou et al. 2020)	61.7	73.1	81.4	82.1	77.6	85.4	89.3	86.1	38.9	63.7	74.7	71.4
MG (Zhou et al. 2021b)	-	-	80.8	-	-	-	87.5	-	-	-	-	-
TransVW (Haghighi et al. 2021)	61.2	69.7	81.2	81.9	78.2	84.5	88.2	86.7	42.7	66.9	76.7	72.3
DiRA (Haghighi et al. 2022)	62.6	74.9	82.7	83.0	78.4	85.2	87.6	87.2	55.3	68.6	83.9	77.8
PCRL (Zhou et al. 2021a)	62.9	75.8	83.0	83.3	78.1	85.5	87.9	87.3	62.1	72.8	81.3	80.4
Adamv1 (Hosseinzadeh Taher, Gotway, and Liang 2023)	60.5	71.6	81.2	82.0	77.9	84.9	87.7	86.2	48.7	67.2	72.8	69.9
Adamv2 (Taher, Gotway, and Liang 2024)	61.6	72.3	82.1	84.1	78.2	85.6	88.6	87.5	54.2	71.9	80.2	76.3
AFiRe (ours)	**63.2**	**76.1**	**83.2**	**84.5**	**78.8**	**86.3**	**89.6**	**89.0**	**68.4**	**76.3**	**92.4**	**87.7**

Table 2: **Transfer learning with different labeling ratios.** AUC ↑ (%) for multi-label classification on NIH and CXP, and Dice ↑ (%) for segmentation on SIIM-ACR are reported. The best results are bolded. '5-C' denotes the five-fold cross-validation.

Model	*Child*CXRs			CXP			NIH		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
MemAE (Gong et al. 2019)	56.5	77.8	82.6	55.6	54.3	53.3	53.3	54.0	50.6
CutPast (Li et al. 2021)	64.0	73.6	72.3	62.7	65.5	60.0	-	-	-
PANDA (Reiss et al. 2021)	65.4	65.7	66.3	66.4	68.6	65.3	55.6	57.4	52.9
M-KD (Salehi et al. 2021)	69.1	74.1	62.3	66.0	69.8	63.6	59.5	61.7	54.7
SALAD (Zhao et al. 2021a)	75.9	82.6	82.1	-	-	-	-	-	-
f-AnoGAN (Schlegl et al. 2019)	74.0	75.5	81.0	63.7	65.8	59.4	62.6	66.5	58.6
SQUID (Xiang et al. 2023)	80.3	87.6	84.7	71.9	78.1	**75.9**	63.7	66.9	59.4
AnatPaste (Sato et al. 2023)	83.0	**91.4**	86.8	70.4	79.2	73.5	63.8	67.4	59.6
AFiRe (ours)	**83.9**	90.8	**87.6**	**72.4**	**79.9**	73.8	**66.8**	**68.2**	**60.4**

Table 3: **Classification results in anomaly detection.** Models are evaluated on ChildCXRs and NIH using official data split and the same settings as SQUID on CXP. ACC scores ↑ (%), AUC scores ↑ (%), F1 scores ↑ (%) are reported.

ods in (iii), we report original AUC scores when available; otherwise, we re-implement them.

AFiRe integrates fine-grained information. We evaluate AFiRe's effectiveness in learning fine-grained representations through anomaly detection experiments. In this experiment, both the pre-trained encoder and decoder are unsupervisedly fine-tuned using Tasks I and II on the target training set. Table 3 presents the results across three datasets, *Child*CXRs, CXP, and NIH, against: (i) recent unsupervised anomaly detection methods for natural images (e.g., MemAE, CutPast, PANDA, M-KD), where AFiRe shows significant improvements (average AUC +15%, +12%, and +10% on *Child*CXRs, CXP, and NIH, respectively), and (ii) state-of-the-art unsupervised methods for medical images (e.g., SALAD, f-AnoGAN, SQUID, AnatPaste), where AFiRe maintains a significant improvement ($p < 0.05$) across all tasks, demonstrating a robust balance between precision and recall. We also present Grad-CAM visualizations to compare the performance of various models, including MAE, DINO, DiRA, PCRL, and the proposed AFiRe, in highlighting lesion regions, as shown in Fig. 7. Compared to the other models, AFiRe demonstrates more precise and focused lesion regions. Especially, in row (e), despite missing some parts of the lesion, AFiRe shows a clear advantage in precision and minimizing false positives, highlighting the

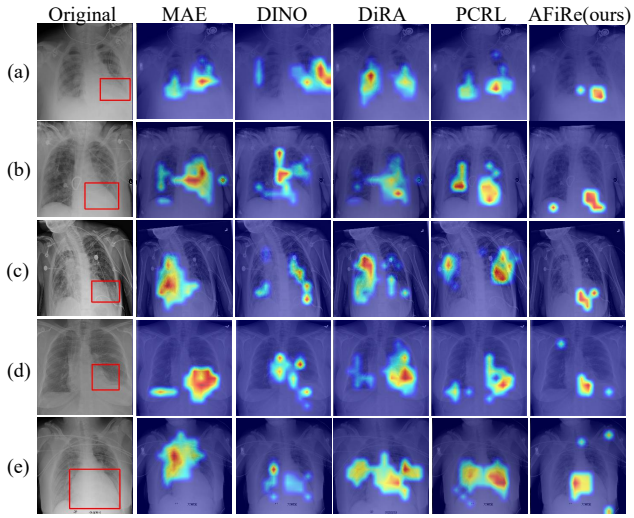


Figure 7: Grad-CAM visualizations on NIH dataset. The lesion regions are annotated by red bounding boxes.

main lesion area more effectively than the other models. *This indicates that AFiRe can better capture and pay attention to specific local anatomical details, effectively distinguishing lesions from surrounding tissues.*

Conclusion

We propose an anatomy-driven self-supervised framework, AFiRe, to incorporate fine-grained anatomical discriminative representations. We synergistically perform this framework with Token-wise anatomy-guided contrastive learning and Pixel-level anomaly-removal restoration. Experimental results on downstream radiographic image diagnosis tasks confirm the robust generalization. Our proposed model exhibits outstanding performance in multi-label classification and disease segmentation. It also provides effective anomaly detection capabilities. However, it tends to prioritize primary abnormalities, potentially overlooking larger affected regions. Our future research is going to improve the model's capability for more accurate detection of lesion regions.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2020YFA0711400, by the National Natural Science Foundation of China under Grant 62171323, U2441252, 62271155, in part by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), in part by the Changjiang Scholars Program of China, in part by the Fundamental Research Funds for the Central Universities.

References

Agu, N. N.; Wu, J. T.; Chao, H.; Lourentzou, I.; Sharma, A.; Moradi, M.; Yan, P.; and Hendler, J. 2021. AnaXNet: anatomy aware multi-label finding classification in chest X-ray. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, 804–813. Springer.

Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Loh, A.; Karthikesalingam, A.; Kornblith, S.; Chen, T.; et al. 2021. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3478–3488.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, H.; Wang, R.; Wang, X.; Li, J.; Fang, Q.; Li, H.; Bai, J.; Peng, Q.; Meng, D.; and Wang, L. 2023. Unsupervised local discrimination for medical images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; and Gelly, S. a. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1705–1714.

Haghighi, F.; Taher, M. R. H.; Gotway, M. B.; and Liang, J. 2022. 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*, 20824–20834.

Haghighi, F.; Taher, M. R. H.; Gotway, M. B.; and Liang, J. 2024. Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? *Medical Image Analysis*, 94: 103086.

Haghighi, F.; Taher, M. R. H.; Zhou, Z.; Gotway, M. B.; and Liang, J. 2021. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10): 2857–2868.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hosseinzadeh Taher, M. R.; Gotway, M. B.; and Liang, J. 2023. Towards foundation models learned from anatomy in medical imaging via self-supervision. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, 94–104. Springer.

Huang, Y.; Lin, L.; Cheng, P.; Lyu, J.; and Tang, X. 2021. Lesion-based contrastive learning for diabetic retinopathy grading from fundus images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 113–123. Springer.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33(01), 590–597.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; Dong, J.; Prasadha, M. K.; Pei, J.; Ting, M. Y.; Zhu, J.; Li, C.; Hewett, S.; Dong, J.; Ziyar, I.; Shi, A.; Zhang, R.; Zheng, L.; Hou, R.; Shi, W.; Fu, X.; Duan, Y.; Huu, V. A.; Wen, C.; Zhang, E. D.; Zhang, C. L.; Li, O.; Wang, X.; Singer, M. A.; Sun, X.; Xu, J.; Tafreshi, A.; Lewis, M. A.; Xia, H.; and Zhang, K. 2018. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5): 1122–1131.e9.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Li, Z.; Yang, L. T.; Ren, B.; Nie, X.; Gao, Z.; Tan, C.; and Li, S. Z. 2024. MLIP: Enhancing Medical Visual Representation with Divergence Encoder and Knowledge-guided Contrastive Learning. *arXiv preprint arXiv:2402.02045*.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6707–6717.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

Reiss, T.; Cohen, N.; Bergman, L.; and Hoshen, Y. 2021. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2806–2814.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14902–14912.

Sato, J.; Suzuki, Y.; Wataya, T.; Nishigaki, D.; Kita, K.; Yamagata, K.; Tomiyama, N.; and Kido, S. 2023. Anatomy-aware self-supervised learning for anomaly detection in chest radiographs. *iScience*.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Langs, G.; and Schmidt-Erfurth, U. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54: 30–44.

Singh, A.; Gorade, V.; and Mishra, D. 2024. MLVICX: Multi-Level Variance-Covariance Exploration for Chest X-ray Self-Supervised Representation Learning. *arXiv preprint arXiv:2403.11504*.

Sowrirajan, H.; Yang, J.; Ng, A. Y.; and Rajpurkar, P. 2021. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, 728–744. PMLR.

Taher, M. R. H.; Gotway, M. B.; and Liang, J. 2024. Representing Part-Whole Hierarchies in Foundation Models by Learning Localizability Composability and Decomposability from Anatomy via Self Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11269–11281.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vu, Y. N. T.; Wang, R.; Balachandar, N.; Liu, C.; Ng, A. Y.; and Rajpurkar, P. 2021. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, 755–769. PMLR.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. 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*, 2097–2106.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xiang, T.; Zhang, Y.; Lu, Y.; Yuille, A. L.; Zhang, C.; Cai, W.; and Zhou, Z. 2023. SQUID: Deep Feature In-Painting for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23890–23901.

Yun, S.; Lee, H.; Kim, J.; and Shin, J. 2022. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8354–8363.

Zhao, H.; Li, Y.; He, N.; Ma, K.; Fang, L.; Li, H.; and Zheng, Y. 2021a. Anomaly detection for medical images using self-supervised and translation-consistent features. *IEEE Transactions on Medical Imaging*, 40(12): 3641–3651.

Zhao, T.; Cao, K.; Yao, J.; Nogues, I.; Lu, L.; Huang, L.; Xiao, J.; Yin, Z.; and Zhang, L. 2021b. 3D graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13743–13752.

Zhou, H.-Y.; Lu, C.; Yang, S.; Han, X.; and Yu, Y. 2021a. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3499–3509.

Zhou, H.-Y.; Yu, S.; Bian, C.; Hu, Y.; Ma, K.; and Zheng, Y. 2020. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 398–407. Springer.

Zhou, Z.; Sodha, V.; Pang, J.; Gotway, M. B.; and Liang, J. 2021b. Models genesis. *Medical image analysis*, 67: 101840.