Exploiting Structural Consistency of Chest Anatomy for Unsupervised Anomaly Detection in Radiography Images

Tiange Xiang*, Yixiao Zhang*, Yongyi Lu, Alan Yuille, *Senior Member, IEEE,* Chaoyi Zhang, Weidong Cai, *Member, IEEE,* Zongwei Zhou, *Member, IEEE*

Abstract—Radiography imaging protocols focus on particular body regions, therefore producing images of great similarity and yielding recurrent anatomical structures across patients. Exploiting this structured information could potentially ease the detection of anomalies from radiography images. To this end, we propose a Simple Space-Aware Memory Matrix for In-painting and Detecting anomalies from radiography images (abbreviated as SimSID). We formulate anomaly detection as an image reconstruction task, consisting of a space-aware memory matrix and an in-painting block in the feature space. During the training, SimSID can taxonomize the ingrained anatomical structures into recurrent visual patterns, and in the inference, it can identify anomalies (unseen/modified visual patterns) from the test image. Our SimSID surpasses the state of the arts in unsupervised anomaly detection by +8.0%, +5.0%, and +9.9% AUC scores on ZhangLab, COVIDx, and CheXpert benchmark datasets, respectively. Code: https://github.com/MrGiovanni/SimSID

Index Terms—Unsupervised Anomaly Detection, Radiography Image Analysis, Image In-Painting.

1 Introduction

Vision tasks in photographic and radiographic images differ significantly. In photographic object identification, the object's location within the image is typically less important—a cat remains a cat regardless of its position within the image. Conversely, in radiography, the relative location and orientation of anatomical structures are crucial for both identifying normal anatomy and recognizing pathologies [1]–[5]. Due to standardized imaging protocols in radiography, images exhibit a high degree of similarity across patients, equipment manufacturers, and institutions (see examples in Figure 1). Consistent and recurrent anatomy can facilitate the analysis of numerous critical problems and should be considered a significant advantage of radiography imaging. For example, several investigations have demonstrated the value of harnessing this prior knowledge to enhance Deep Nets' performance, such as adding location features, modifying objective functions, and constraining coordinates relative to landmarks in images [6]–[12]. This paper focuses on unsupervised anomaly detection, seeking to answer the critical question: *Can we exploit consistent anatomical patterns and their spatial information to strengthen Deep Nets in detecting anomalies from radiography images without manual annotation?*

Unsupervised anomaly detection only uses healthy images for model training and requires no other annotations such as disease diagnosis or localization [13]. As many as 80% of clinical errors occur when the radiologist misses the abnormality in the first place [14]. The clinical impact of anomaly detection is to reduce that 80% by clearly pointing out to radiologists that there exists a suspicious lesion and then having them look at the scan in depth. We formulate the task of anomaly detection as an in-painting task to exploit the anatomical consistency in appearance, position, and layout in the chest region. Specifically, we propose a Simple Space-Aware Memory Matrix for In-painting and Detecting anomalies from radiography images (abbreviated as SimSID). In the training phase, our model can *dynamically* maintain a visual pattern dictionary by taxonomizing the recurrent anatomical structure based on its spatial locations. Due to the consistency in anatomy, the same body region across normal images is expected to express similar visual

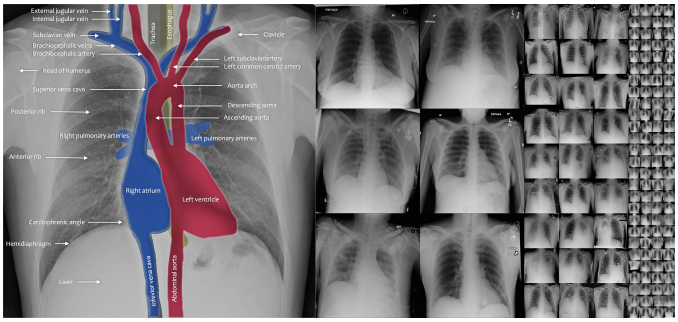


Fig. 1. Anomaly detection in radiography images can be both easier and harder than photographic images. It is easier because radiography images are spatially structured due to consistent imaging protocols. It is harder because anomalies are subtle and require medical expertise to annotate. We contribute a novel anomaly detection method (SimSID) that directly exploits the structured information in radiography images.

- *Tiange Xiang and Yixiao Zhang contribute equally.
- Corresponding author: Zongwei Zhou (ZZHOU82@JH.EDU).
- T. Xiang, C. Zhang, W. Cai are with the School of Computer Science, University of Sydney, Camperdown NSW 2006, Australia. {txia7609@uni,czha5168,tom.cai}.sydney.edu.au; lchen025@e.ntu.edu.sg
- Y. Zhang, Y. Lu, A. Yuille, Z. Zhou are with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA. {yzhan334,zzhou82}@jh.edu; {yylu1989,alan.l.yuille}@gmail.com

patterns, which makes the total number of unique patterns manageable. In the inference, since anomaly patterns are not present in the learned dictionary, the reconstructed image is expected to be unrealistic. As a result, the model can identify the anomaly by assessing the quality of the in-painting task. The success of anomaly detection has two basic assumptions [15]: *first*, anomalies only occur very rarely in the training data (or a small proportion in an image); *second*, anomalies differ from the normal patterns significantly. Consequently, the learned dictionary will reflect the general distribution of anatomical patterns in normal human anatomy. Notably, our SimSID is robust to a level of abnormal images in the training set by automatically omitting minority anatomical patterns (evidenced in Figure 9). This should be considered a significant advantage because it can largely relax the requirement of disease-free images for training existing unsupervised anomaly detection methods [13].

We have conducted extensive experiments on *three* large-scale radiography imaging datasets. Our SimSID is significantly superior to 21 predominant methods in unsupervised anomaly detection by at least 8.0% on the ZhangLab dataset, yielding an AUC of 91.1%; 5.0% AUC gain on the COVIDx dataset, with an AUC of 83.5%; additionally, we have demonstrated a 9.9% improvement over the state of the arts on the Stanford CheXpert dataset, with an AUC of 79.7%. The quantitative results and qualitative visualization show the superiority of SimSID over the state of the arts.

2 Related Work

2.1 Anomaly Detection in Natural Imaging

Anomaly detection is the task of identifying rare events that deviate from the distribution of normal data [16]. Early attempts include one-class SVM [17], dictionary learning [18], and sparse coding [19]. Due to the lack of sufficient samples of anomalies, later works typically formulate anomaly detection as an unsupervised learning problem [20]–[29]. These can be roughly categorized into reconstruction-based and density-based methods. Reconstruction-based methods train a model (e.g., Auto-Encoder) to recover the original inputs [30]–[35]. The anomalies are identified by subtracting the reconstructed image from the input image. Density-based methods predict anomalies by estimating the normal data distribution (e.g., via VAEs [36] or GANs [37], [38]). However, their learned distribution for normal images cannot explain possible abnormalities. In this paper, we address these limitations by maintaining a visual pattern dictionary which is extracted from homogeneous medical images.

Several other previous works investigated the use of image in-painting for anomaly detection, i.e., parts of the input image are masked out and the model is trained to recover the missing parts in a self-supervised way [39]–[43]. There are also plenty of works on detecting anomalies in video sequences [44]–[47]. Recently, Bergmann et al. [48] and Salehi et al. [49] proposed student-teacher networks similar to ours, whereas our method utilizes such a structure to distillate input-aware features only, and the teacher network is completely disabled during inference.

2.2 Anomaly Detection in Medical Imaging

Anomaly detection in the medical domain is usually approached at per pathology-basis [50]–[54]. There are *supervised* anomaly detection methods to identify specific types of abnormalities, such as vascular lesions [55], malignant melanoma [56], brain tumors [57], [58], and pulmonary nodules/embolism [59], [60]. Recent *unsupervised* anomaly detection methods have been proposed to detect anomalies in general [13], [61]–[64]. With the help of GANs, anomaly detection can be achieved with *weak* annotation. In AnoGAN [65], the discriminator was heavily over-fitted to the normal image distribution to detect the anomaly. Subsequently, f-AnoGAN [38] was proposed to improve computational efficiency. Naval et al. [66] designed an autoencoder network to fit the distribution of normal images. The spatial coordinates and anomaly probabilities are mapped over a proxy for different tissue types. Han et al. [67] proposed a two-step GAN-based framework for detecting anomalies in MRI slices as well. However, their method relies on a voxel-wise representation for the 3D MRI sequences, which is impossible in our task. Most recently, a hybrid framework SALAD [68] was proposed that combines GAN with self-supervised techniques. Normal images are first augmented to carry the forged anomaly through pixel corruption and pixel shuffling. The fake abnormal images, along with the original normal ones, are fed to the GAN for learning more robust feature representations. However, these approaches demand strong prior knowledge and assumptions about the anomaly type to make the augmentation effective.

Incorporating memory modules into neural networks has been demonstrated to be effective for many tasks [69]–[73]. Adopting a Memory Matrix for unsupervised anomaly detection was first proposed in MemAE [74]. In addition to auto-encoding (AE), Gong et al. injected an extra Memory Matrix between the encoder and the decoder to capture normal feature patterns during training. The matrix is jointly optimized along with the AE and hence learns an essential basis to be able to assemble normal patterns. Based on this paradigm, Park et al. [75] introduced a non-learnable memory module that can be updated with inputs. Considering the extra memory usage in existing methods, Lv et al. [76] proposed a dynamic prototype unit that encodes normal dynamics on the fly, while consuming little additional memory.

With the recent progress in diffusion models [77]–[80], it is also feasible to achieve anomaly detection by generating normal image samples. One of the earliest attempts that followed this paradigm for medical anomaly detection was proposed by Wolleb et al. [81]. SynDiff [82], as one of the following-up methods, extended the generation quality further by incorporating adversarial projections during the inverse diffusion process.

Differing from photographic images, radiography imaging protocols produce images with consistent anatomical patterns, and meanwhile, the anomalies in radiography images can be subtle in appearance and hard to interpret (Figure 1). Unlike most existing works, we present a novel method that explicitly harnesses the radiography images' properties, therefore dramatically improving the performance in anomaly detection from radiography images.

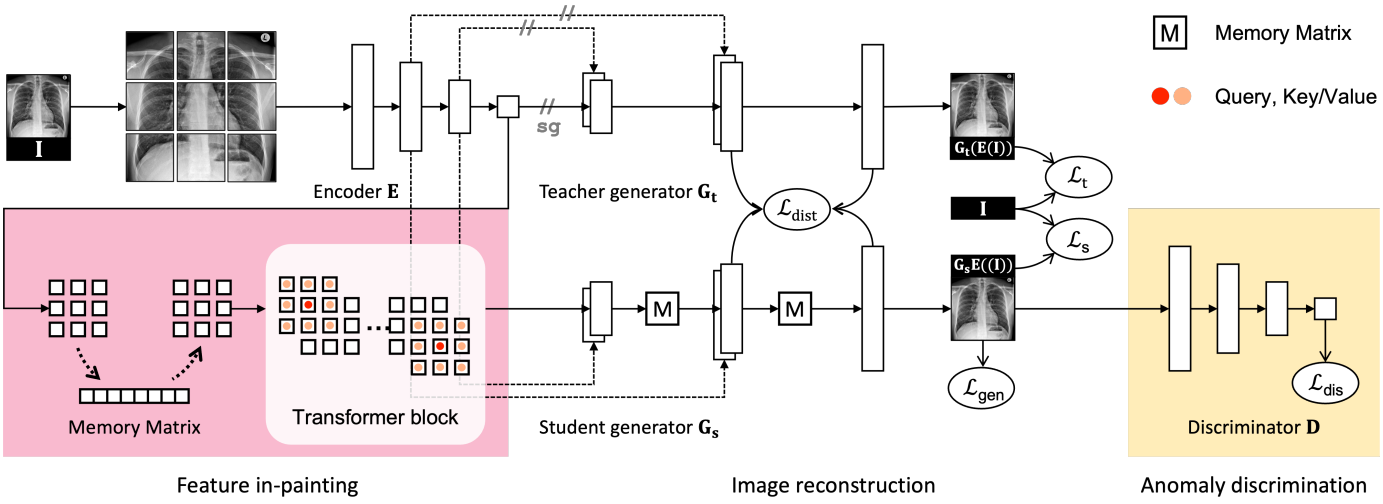


Fig. 2. **SimSID overview.** We divide an input image into $N \times N$ non-overlapping patches and feed them into the encoder for feature extraction. Two generators will be trained to reconstruct the original image. Along with the reconstruction, a dictionary of anatomical patterns will be created and updated dynamically via a novel space-aware memory matrix (§3.2); The teacher generator directly uses the features extracted by the encoder; the student generator uses the features augmented by a new feature in-painting block (§3.3). The teacher and student generators are coupled through a knowledge distillation paradigm. We employ a discriminator to assess whether the image reconstructed by the student generator is real or fake. Once trained, the discriminator can be used to detect anomalies in test images (§3.4).

2.3 Our Previous Work

We first presented Space-aware Memory Queues for In-painting and Detecting anomalies from radiography images (SQUID) published in CVPR-2023 [83]. This paper is a significant extension with the following four improvements.

1) We have introduced new notations, formulas, and diagrams, as well as detailed methodology descriptions along with their learning objectives, for a succinct framework overview.
2) We have significantly simplified the framework by removing the Memory Queue and masked shortcut while achieving higher performance and easing the training than SQUID [83].
3) We have examined SimSID with 21 existing unsupervised (and also one weakly-supervised) anomaly detection methods on three radiography imaging tasks, showing that SimSID surpasses all these methods by a large margin, as well as SQUID [83].
4) We have investigated the robustness of our SimSID to the normal/abnormal ratio in the training set, relaxing the requirement of the disease-free training set of existing anomaly detection approaches.

3 SimSID

3.1 Overview

Feature extraction: We divide the input image into $N \times N$ non-overlapping patches, then use a CNN encoder to extract features for each patch. The extracted features will be used for image reconstruction. Practically, the encoder can be any backbone architectures, and for simplicity, we adopt basic Convolutions and Pooling layers in the experiments.

Feature in-painting: A dictionary of normal anatomical patterns will be created and updated dynamically through a Memory Matrix (§3.2). The extracted patch features will be substituted by the most close items in the matrix. Then, the substituted features of each image patch would be masked out, a transformer block (§3.3) is used to predict the masked feature based on the surrounding patch features.

Image reconstruction: We introduce teacher and student generators to reconstruct the original image. Specifically, the teacher generator reconstructs the image using the features extracted by the encoder directly (essentially an auto-encoder [84]). The student generator, on the other hand, reconstructs the image using the features augmented by our in-painting block. The teacher and student generators are coupled through knowledge distillation [85] at all the up-sampling levels. The objective of the student generator is to reconstruct a normal image from the augmented features; the reconstructed image will then be used for anomaly discrimination (§3.4); while the teacher generator[1] serves as a regularizer to prevent the student generator from collapsing[2] (constantly generating the same normal image).

Anomaly discrimination: Following the adversarial learning [38], [65], we employ a discriminator to assess whether the generated image is real or fake. Both teacher and student generators will receive the gradient derived from the discriminator. The two generators and the discriminator are competing against each other in a way that, together, they converge to an equilibrium. Once trained, the discriminator can be used to detect anomalies in test images (§3.4).

3.2 Developing Space-aware and Hierarchical Memory

Motivation: The Memory Matrix was initially introduced by Gong et al. [74] and has since been widely adopted in unsupervised anomaly detection [46], [87]–[90]. To forge a "normal" appearance, the features are *augmented* by weighted

1. We disabled the backpropagation between the teacher generator and encoder by stop-gradient [86] and showed its benefit in Table 2.
2. Alternative strategies to avoid collapse include early stopping, elastic regularization, or Gaussian prior [7], [39].

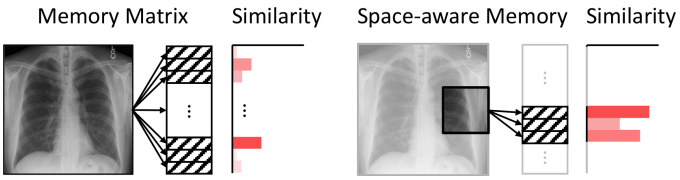


Fig. 3. **Space-aware memory.** For unique encoding of location information, we restrict each patch to be only able to access a set of specific tokens in the memory.

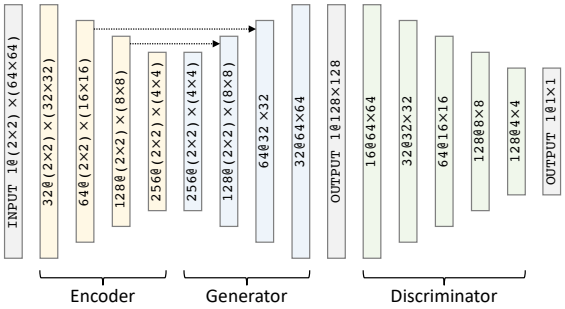


Fig. 4. **SimSID architecture.** Our SimSID consists of an encoder, a student generator, a teacher generator, and a discriminator. All of the network architectures are built with plain convolution, batch normalization, and ReLU activation layers. Given an input image, we first divide it into non-overlapping patches. The encoder then extracts the patch features. The student and teacher generators were constructed identically. The only difference is that additional Memory Matrices are placed in the student generator. The discriminator was constructed in a more lightweight style. Note that the images are discriminated at their full resolution rather than in patches.

averaging the similar patterns in Memory Matrix. This augmentation is, however, applied to the features extracted from the whole image, discarding the location and spatial information embedded in images. Therefore, Memory Matrix in its current form cannot perceive the anatomical consistency that radiography images can offer.

Space-aware memory: To harness the spatial information, we pass the divided patches into the model rather than the entire image. These patches are associated with unique location information of the original image. We seek to build the relationship between the patch location and memory region. The memory matrix $\mathbf{M}$ is divided into blocks $\{\mathbf{M}_{i,j} \in \mathbb{R}^{N\times C}\}$, each associated with a patch at location (i, j), where N and C denote the number and the dimension of items, respectively. Let $\mathbf{z}_{i,j} \in \mathbb{R}^{C}$ denote the feature of patch (i, j), we obtain the augmented feature $\hat{\mathbf{z}}$ as follows:

$$\hat{\mathbf{z}}_{\mathbf{i,j}} = \sum_{k=1}^{N} G(s^k)\mathbf{M}_{i,j}^k, \tag{1}$$

where s^k is the similarity score computed by dot product between the patch feature $\mathbf{z}_{i,j}$ and the k-th memory item $\mathbf{M}_{i,j}^k$. $G(\cdot)$ is the Gumbel-softmax operation, which shrinks the number of activated memory items[3]. With the division of image patches together with memory blocks, a patch derived from a particular location can only search for similar items within a specific block in the Memory Matrix (illustrated in Figure 3). We refer to this new searching strategy as "space-aware memory". This strategy can also accelerate the searching speed compared with [74] as it no longer has to go through the entire Memory Matrix to assemble similar features. Results in Table 2 highlight the significance of space-aware memory (AUC improved from 77.6% to 91.1%).

Hierarchical memory: The use of one memory matrix at the deepest layer in the encoder is insufficient to reconstruct high-quality image with details. To capture anatomical patterns at different scales, we placed a space-aware memory matrix at several levels of the generator to create a hierarchy of scales. Studies in [74] discovered that too many memories can lead to excessive information filtering and degrade the model's capacity to retain the most representative normal patterns instead of all needed ones. This problem is solved by adding skip connections between the encoder and the generator[4]. In each generator layer, the feature map is upsampled and concatenated with low-level features carried by the skip connection, then filtered by the following space-aware Memory Matrix. We empirically found that a total of three Memory Matrices (one in the feature in-painting block and two in the generator) are sufficient. This design is also proved to be effective by [46] in flow-guided video anomaly detection. Results in Table 2 highlight the significance of hierarchical memory and skip connection, achieving an AUC improvement of 8.2% and 11.6%.

3.3 In-painting Features by Learned Memory Matrix

Motivation: Image in-painting [93] was initially proposed to recover corrupted regions in the image based on the available neighboring context. The recovered regions, however, have been seen to associate with boundary artifacts, distorted and blurry predictions, particularly when using methods based on Deep Nets [41], [94]. These undesired artifacts are responsible for numerous false positives when formulating anomaly detection as an image in-painting task [31], [33]. It is because the subtraction between input and output will reveal artifacts generated by Deep Nets instead of true anomalies. To alleviate this issue, we propose the in-painting task at the feature level rather than the image pixel level. Latent features are invariant to subtle noise, rotation, and translation in the pixel level and therefore are expected to be more suitable for anomaly detection. The model predicts central features based on neighboring features. This in-painting step is repeated for all of the patch neighborhoods through sliding-window with stride of 1. Similar to the sliding-window as in convolutions, the whole process is fully parallelizable and computed efficiently.

In-painting block: We integrate our Memory Matrix with a novel in-painting block to perform an in-painting task at

3. Controlling the number of activated memory items has proven to be advantageous for anomaly detection [91]. However, setting a hard shrinkage threshold as in [74] fails to adapt to cases where abnormal signals are sufficient to reconstruct a normal image. Inspired by [92], we present a *Gumbel Shrinkage* schema: only activating the top-k most similar memory items during the forward pass and distributing the gradient to all patterns during back-propagation. Gumbel Shrinkage improves AUC from 86.2% to 91.1% (see Table 2).

4. It is worth noting that the outermost skip connection should not be added (shown in Figure 4). It is because a memory matrix must be followed by skip connections; otherwise, the reconstruction might be fulfilled by the highest-level encoding-decoding information, making all other lower-level encoding, decoding and memory blocks not work.

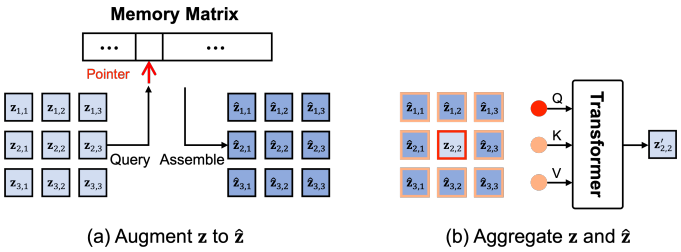


Fig. 5. **Two-step workflow of the in-painting block.** (a) Each non-overlapping patch feature $\mathbf{z}$ is queried to an unique region in Memory Matrix, the most similar items are assembled to $\hat{\mathbf{z}}$. (b) Each center patch feature $\mathbf{z}$ and its eight neighbors $\hat{\mathbf{z}}$ are used as query and key/value respectively to a Transformer layer for in-painting. During training, the Memory Matrix is updated through optimization via backpropagation.

the feature level. The $w \times h$ non-overlapping patch features $\mathbf{z}_{\{(1,1),\cdots,(w,h)\}}$ are augmented to the most similar "normal" patterns $\hat{\mathbf{z}}_{\{(1,1),\cdots,(w,h)\}}$ in Memory Matrix (Figure 5a). Since $\hat{\mathbf{z}}$ is assembled by patterns from previously seen images, it is not subject to the current input image. To recap characteristics of the current image, naturally, we aggregate both patch features $\mathbf{z}$ and augmented features $\hat{\mathbf{z}}$ using a Transformer block [95]. For each patch $\mathbf{z}_{i,j}$, its spatially adjacent eight "normal" patches $\hat{\mathbf{z}}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}}$ are used as conditions to refine $\mathbf{z}_{i,j}$ (Figure 5b). The query token is flattened $\mathbf{z}_{(i,j)} \in \mathbb{R}^{1\times *}$ and key/value tokens are $\hat{\mathbf{z}}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}} \in \mathbb{R}^{8\times *}$. At the start and the end of our in-painting block, we apply an extra pair of point-wise convolutions (1×1 convolutional kernel) to reduce feature dimensions and accelerate the training process.

3.4 Anomaly Discrimination

Our in-painting block focuses on augmenting any patch feature (either normal or abnormal) into the normal feature pattern. The student generator will then reconstruct a "normal" image based on the augmented features. The teacher generator is used to preserve the normal image intact and prevent the student generator from collapsing. Once trained, the semantic (rather than pixel-level) difference between the input and the reconstructed image is expected to be small if normal; the semantic difference will be big if there are anomalies. We therefore delegate the optimized discriminator network for alerting anomalies perceptually. Unlike common approaches that use pixel-level comparisons to alert anomaly [68], we are trying to utilize a discriminator to assess the generation of 'normal samples'. This allows SimSID to be more robust to pixel-level noise and variations. For better clarification, we notate the encoder, teacher generator, student generator, and discriminator as $\mathbf{E}$, $\mathbf{G}_\text{t}$, $\mathbf{G}_\text{s}$, and $\mathbf{D}$. An anomaly score (A) can be computed through:

$$A = \phi\left(\frac{\mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I}))) - \mu}{\sigma}\right), \tag{2}$$

where $\phi(\cdot)$ is the Sigmoid function, μ and σ are the mean and standard deviation of anomaly scores calculated on all training samples.

3.5 Loss Function

Our SimSID is optimized by five loss functions. The mean square error (MSE) between input and reconstructed images is used for both teacher and student generators. Concretely, for the teacher and student generators, we have:

$$\mathcal{L}_\text{t} = ||\mathbf{I} - \mathbf{G}_\text{t}(\mathbf{E}(\mathbf{I}))||^2, \quad \mathcal{L}_\text{s} = ||\mathbf{I} - \mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I}))||^2, \tag{3}$$

where $\mathbf{I}$ denotes the input image. Following the knowledge distillation paradigm, we apply a distance constraint between the teacher and student generators at all levels:

$$\mathcal{L}_\text{dist} = \sum_{i=1}^{l} (\mathbf{z}_\text{t}^i - \mathbf{z}_\text{s}^i)^2, \tag{4}$$

where l is the level of features used for knowledge distillation, $\mathbf{z}_\text{t}$ and $\mathbf{z}_\text{s}$ are the intermediate features in the teacher and student generators, respectively. In addition, we employ an adversarial loss (similar to DCGAN [96]) to improve the quality of the image generated by the student generator. Specifically, the following equation is minimized:

$$\mathcal{L}_\text{gen} = \log(1 - \mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I})))). \tag{5}$$

The discriminator seeks to maximize the probability for real images and the inverted probability for fake images:

$$\mathcal{L}_\text{dis} = \log(\mathbf{D}(\mathbf{I})) + \log(1 - \mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I})))). \tag{6}$$

In summary, our SimSID is trained to *minimize* the generative loss terms ($\lambda_\text{t}\mathcal{L}_\text{t} + \lambda_\text{s}\mathcal{L}_\text{s} + \lambda_\text{dist}\mathcal{L}_\text{dist} + \lambda_\text{gen}\mathcal{L}_\text{gen}$) and to *maximize* the discriminative loss term ($\lambda_\text{dis}\mathcal{L}_\text{dis}$).

4 Experiments

4.1 Public Chest Radiography Benchmarks

ZhangLab Chest X-ray [109]: This dataset contains healthy and pneumonia images, *officially* split into training and test sets. The training set consists of 1,349 normal and 3,883 abnormal images; the test set has 234 normal and 390 abnormal images. We randomly separate 200 images (100 normal and 100 abnormal) from the training set as the validation set for early-stopping. Since the images are of varying sizes, we resized all the images to 128×128. We used this dataset for ablation studies as well.

Stanford CheXpert [110]: We conducted evaluations on the front-view PA images in the CheXpert dataset, which account for a total of 12 different anomalies. In all front-view PA scans, there are 5,249 normal and 23,671 abnormal images for training; 250 normal and 250 abnormal images (with at least 10 images per disease type) from the training set for testing; 14 normal and 19 abnormal images for early-stopping (`val` set based on the *official* split). All images are resized to 128×128 as inputs.

COVIDx [111]: The original dataset contains a train and a test set. The train set has 29,187 chest radiographs, of which 8,085 are normal, 5,555 are non-covid pneumonia and 15,547 are COVID-19 positive. The test set has 400 chest X-rays, of which 100 are normal, 100 are non-covid pneumonia and the rest 200 are COVID-19 positive. We randomly separate 400 images (200 normal, 100 non-covid pneumonia and 100 COVID-19 pneumonia) from the training set as the validation set. COVIDx v9 was used in our experiments.

TABLE 1
Benchmark results on the *official* test sets of the three datasets. Apart from those performances directly taken from other literature, we present the mean and standard deviation (mean±s.d.) across three different trials for all models. For every dataset, the AUC improvement between our SimSID and the best alternative baseline method is significant at $p = 0.05$ level, performed by an independent two sample t-test.

Dataset: *ZhangLab*	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Auto-encoder†	-	59.9	63.4	77.2
VAE† [36]	Arxiv'13	61.8	64.0	77.4
Ganomaly† [37]	ACCV'18	78.0	70.0	79.0
f-AnoGAN† [38]	MIA'19	75.5	74.0	81.0
MemAE [74]	ICCV'19	77.8±1.4	56.5±1.1	82.6±0.9
Fixed-Point GAN‡ [31]	ICCV'19	83.1	78.0	84.3
MNAD [75]	CVPR'20	77.3±0.9	73.6±0.7	79.3±1.1
SALAD† [68]	TMI'21	82.7±0.8	75.9±0.9	82.1±0.3
CutPaste [97]	CVPR'21	73.6±3.9	64.0±6.5	72.3±8.9
PANDA [39]	CVPR'21	65.7±1.3	65.4±1.9	66.3±1.2
M-KD [49]	CVPR'21	74.1±2.6	69.1±0.2	62.3±8.4
IF 2D [66]	MICCAI'21	81.0±2.8	76.4±0.2	82.2±2.7
PaDiM [98]	ICPR'21	71.4±3.4	72.9±2.4	80.7±1.2
IGD [99]	AAAI'22	73.4±1.9	74.0±2.2	80.9±1.3
SQUID [83]	Ours (CVPR'23)	87.6±1.5	80.3±1.3	84.7±0.8
SimSID	Ours	**91.1±0.9**	**85.0±1.0**	**88.0±1.1**

†The results are taken from Zhao et al. [68]; ‡Fixed-Point GAN is considered as a baseline of weakly supervised learning (requiring image-level labels)

Dataset: *COVIDx*	Ref & Year	AUC (%)	Acc (%)	F1 (%)
DAE* [100]	ICANN'11	55.7		
ALAD† [101]	ICDM'18	58.0		
Ganomaly† [37]	ACCV'18	58.4		
OCGAN* [102]	CVPR'18	61.2		
f-AnoGAN‡ [38]	MIA'19	66.9		
MemAE [74]	ICCV'19	71.8±3.6	77.1±2.1	86.4±0.8
ADGAN* [103]	ISBI'19	65.9		
CCD+IGD* [104]	MICCAI'21	74.6		
PaDim* [98]	ICPR'21	61.4		
PatchCore† [105]	Arxiv'21	52.0		
CutPaste [97]	CVPR'21	78.5±2.3	**83.1±0.4**	**89.5±0.2**
PANDA [39]	CVPR'21	72.3±1.0	76.9±0.8	86.4±0.4
M-KD [49]	CVPR'21	71.7±1.1	69.7±4.5	55.6±2.5
MS-SSIM* [99]	AAAI'22	63.4		
IGD* [99]	AAAI'22	69.9		
SQUID [83]	Ours (CVPR'23)	74.7±0.9	76.8±0.1	86.0±0.2
SimSID	Ours	**83.5±0.6**	82.6±0.6	88.8±0.1

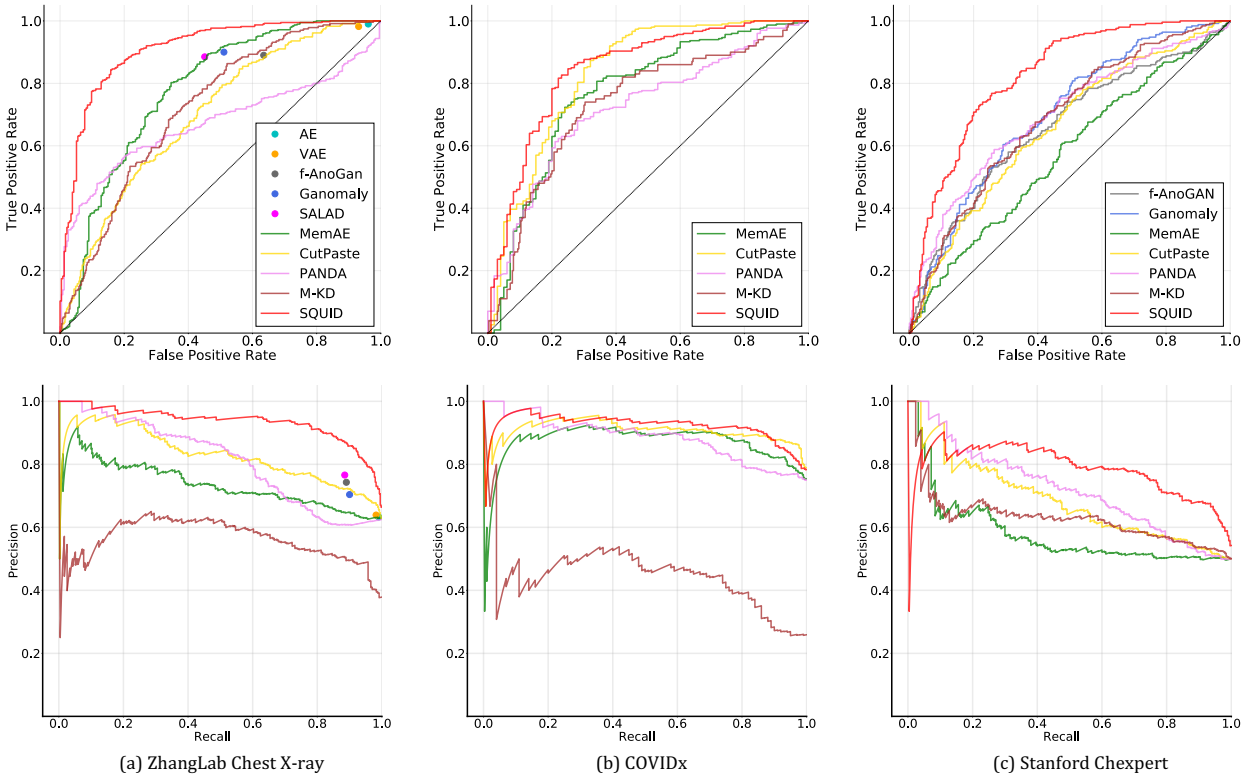
*The results are taken from Tian et al. [106]; †The results are taken from Rahman Siddiquee et al. [107]; ‡The results are taken from Tian et al. [108]

Dataset: *CheXpert*	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Ganomaly [37]	ACCV'18	68.9±1.4	65.7±0.2	65.1±1.9
f-AnoGAN [38]	MIA'19	65.8±3.3	63.7±1.8	59.4±3.8
MemAE [74]	ICCV'19	54.3±4.0	55.6±1.4	53.3±7.0
CutPaste [97]	CVPR'21	65.5±2.2	62.7±2.0	60.3±4.6
PANDA [39]	CVPR'21	68.6±0.9	66.4±2.8	65.3±1.5
M-KD [49]	CVPR'21	69.8±1.6	66.0±2.5	63.6±5.7
SQUID [83]	Ours (CVPR'23)	78.1±5.1	71.9±3.8	**75.9±5.7**
SimSID	Ours	**79.7±2.2**	**72.9±1.9**	71.9±2.3

4.2 Baselines, Metrics, and Implementation

We considered a total of 21 major baselines for direct comparison (elaborated in Table 1): for example, Auto-encoder, VAE [36]—the classic UAD methods; Ganomaly [37], f-AnoGAN [38], IF [66], SALAD [68]—the current state of the arts for medical imaging; and MemAE [74], CutPaste [97], M-KD [49], PANDA [39], PaDiM [98], IGD [112]—the most recent UAD methods. We evaluated performance using standard metrics: receiver operating characteristic (ROC) curve, precision-recall (PR) curves, area under the ROC curve (AUC), accuracy (Acc) and F1-score (F1). All results were based on at least *three* independent runs.

We utilized common data augmentation strategies such as random translation within the range $[-0.05, +0.05]$ in four directions and a random scaling within the range of $[0.95, 1.05]$. The Adam optimizer was used with a batch size of 16 and a weight decay of $1e$-5. The learning rate was initially set to $1e$-4 for both the generator and the discriminator and then decayed to $2e$-5 in 200 epochs following the cosine annealing scheduler. The discriminator was trained at every iteration, while the generator was trained every two iterations. We set the loss weights as $\lambda_{\text{t}} = 0.01$, $\lambda_{\text{s}} = 10$, $\lambda_{\text{dist}} = 0.001$, $\lambda_{\text{gen}} = 0.005$, and $\lambda_{\text{dis}} = 0.005$. We divided the input images in 4×4 non-overlapping patches generator. The architectures of our generators and discriminator are detailed in Figure 4.



(a) ZhangLab Chest X-ray (b) COVIDx (c) Stanford Chexpert

Fig. 6. ROC curves and PR space comparison on the ZhangLab Chest X-ray, COVIDx and Stanford CheXpert datasets. ROC = receiver operating characteristic; PR = precision-recall.

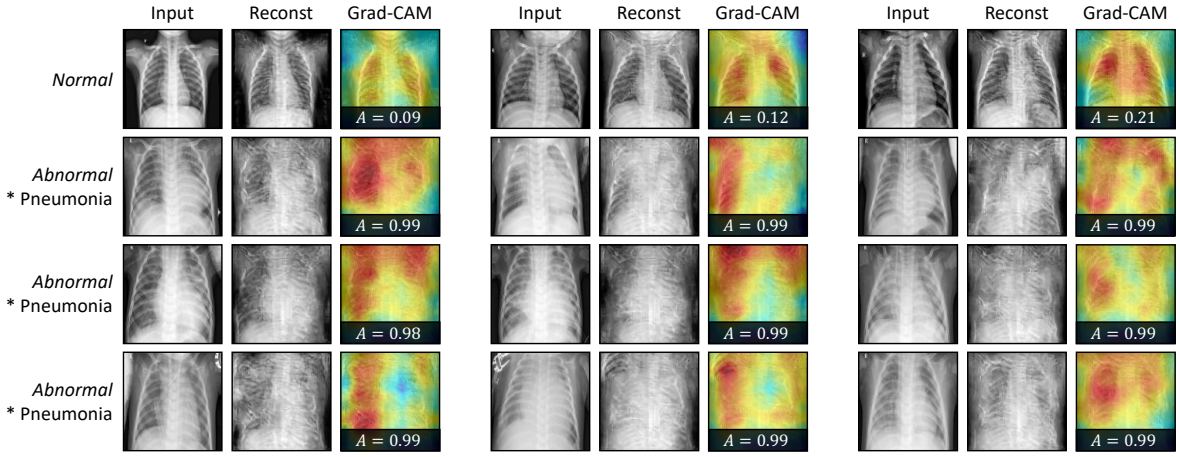


Fig. 7. [Better viewed on-line, in color, and zoomed in for details] Reconstruction results of SimSID on the ZhangLab dataset. The corresponding Grad-CAM heatmaps, along with anomaly scores, are shown. The anomaly score denotes the probability of the image containing abnormal.

5 Results

5.1 Benchmarking SimSID on Three Public Datasets

Our SimSID was mainly evaluated on three large-scale benchmarks: ZhangLab Chest X-ray, COVIDx and Stanford CheXpert for comparing with a wide range of state-of-the-art methods. According to Table 1, our SimSID achieves the most promising result in terms of most metrics on these datasets. Specifically, SimSID outperforms the second best runner-up counterpart SALAD [68] by 8.4% in AUC, 9.1% in Accuracy, and 5.9% in F1 on the ZhangLab dataset. In particular, our SimSID trained in an unsupervised manner surpasses Fixed-Point GAN [31]—a weakly supervised anomaly detection method—by 8% in AUC. Additionally, CutPaste [97] and M-KD [49] were previous state of the arts on COVIDx and CheXpert datasets, respectively. Our SimSID not only achieves 5.0% and 9.9% improvements, but also significantly exceeds its previous version (SQUID). The ROC curve and PR curve are presented in Figure 6, demonstrating that our method yields the best trade-off between sensitivity and specificity. Overall, the significant improvements observed with SimSID proved the effectiveness of our proposed designs and techniques in this work.

In Figure 7, we visualize the reconstructions of Sim-

TABLE 2
Component studies indicate that the overall performance benefits from all the components in SimSID. The ablation study is conducted on the Zhanglab dataset.

Method	AUC(%)	Acc(%)	F1(%)
w/o Space-aware Memory	77.6±0.5	75.5±0.5	82.5±0.6
w/o Skip Connection	79.5±1.6	73.0±1.4	78.8±0.5
w/o In-painting Block	80.9±2.1	75.8±1.5	81.6±1.3
w/o Hierarchical Memory	82.9±1.2	77.4±1.1	81.2±0.5
w/o Knowledge Distillation	85.4±0.8	79.5±0.7	83.5±0.8
w/o Stop Gradient	85.0±4.3	77.6±2.8	79.8±1.6
w/o Gumbel Shrinkage	86.2±3.3	80.5±3.2	85.4±2.1
w/ Memory Queue instead	86.7±2.1	80.6±1.7	84.2±1.3
Convolution Layers	86.3±3.4	80.8±3.0	85.4±2.2
Pixel-level In-painting	79.1±0.4	74.4±1.6	81.3±0.9
SQUID [83]	87.6±1.5	80.3±1.3	84.7±0.8
Full SimSID	**91.1±0.9**	**85.0±1.0**	**88.0±1.1**

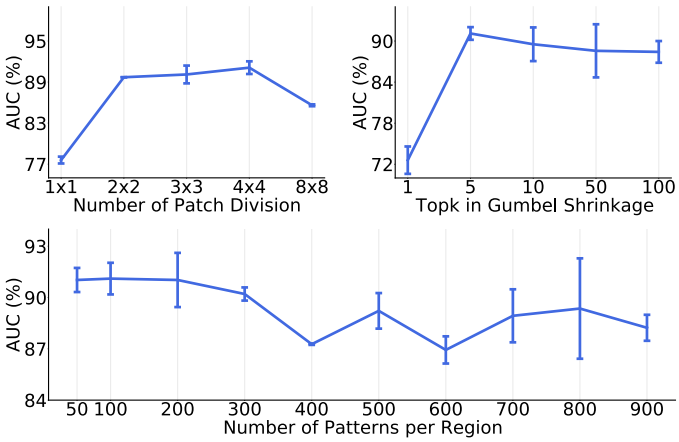


Fig. 8. SimSID is robust to hyper-parameter modifications to some extent. The best result is obtained by dividing 4×4 patches, setting 100 patterns per memory region, and activating the top 5 patterns through Gumbel Shrinkage. The hyper-parameters were determined on the validation set of ZhangLab and were applied to all three datasets.

SID on exemplary normal and abnormal images in the ZhangLab dataset. For normal cases, SimSID can easily find a similar match in the memory and hence achieves the reconstruction smoothly. For abnormal cases, the contradiction will arise by imposing forged normal patterns into the abnormal features. In this way, the generated images will vary significantly from the input, which will then be captured by the discriminator. We plot the heatmap of the discriminator (using Grad-CAM [113]) to indicate the regions that are poorly reconstructed. As a result, the reconstructed healthy images yield much lower anomaly scores than the diseased ones, validating the effectiveness of SimSID.

We also benchmarked the running speed of models to compare the efficiency of SQUID and the proposed SimSID. It is observed that one step training of SimSID is 17.2s faster than SQUID (11.0s *v.s.* 28.2s) and one step inference of SimSID is 0.5s faster than SQUID (9.0s *v.s.* 9.5s).

Limitation: We found SimSID in its current form, is not able to *localize* anomalies at the pixel level precisely. It is understandable because, unlike [49], [114]–[117], our SimSID is an unsupervised method, requiring zero manual annotation for normal/abnormal images. More investigation on pixel-level localization (or even segmentation) and multi-scale detections could be meaningful in the future.

5.2 Ablating Key Properties in SimSID

Component study: We first examine the impact of components in SimSID by taking each one of them out of the entire framework. Table 2 shows that each component accounts for at least 5% performance gains. The space-aware memory (+13.5%) and in-painting block (+10.2%) are among the most significant contributors, which underline our motivation and justification of the method development (§3.2 and §3.3). Moreover, the knowledge distillation from teacher to student generators strikes an important balance: the student generator reconstructs faithful "normal" images from similar anatomical patterns in the dictionary while preserving the unique characteristics of each input image (regularized by the teacher generator). Besides, we must acknowledge that the training tricks (e.g., hard shrinkage [92], stop gradient [86]) are necessary for the remarkable performance.

We further ablate the feature in-painting design in our model by comparing to other reasonable module designs: In our proposed in-painting block, a transformer layer is used to aggregate the patch features and the Memory Matrix augmented "normal" features. However, one may wonder if a simple convolution layer can also suffice. We conducted experiments by replacing the transformer layer with a convolution layer while preserving other structures. The result is presented in the 8th line of Table 2, where with convolution layer the AUC decreased by 4.8%. Another comparable design is pixel-level in-painting. As discussed in §3.3, raw images usually contain larger noise and artifacts than features, so we proposed to achieve the in-painting at the feature level rather than at the pixel level [33], [93], [118], [119]. To validate our claim, we have conducted experiments on carrying out the in-painting at the pixel level. Instead of using a transformer layer to in-paint the extracted patch features, we randomly zeroed out parts of the input patches with 25% probability and let SimSID in-paint the distorted input images. All other settings and objective functions remain unchanged. The result is shown in the 10th row of Table 2, and feature-level in-painting surpasses pixel-level in-painting by 12.0% in AUC. As shown from the table, we validate that the new space-aware memory matrix, the in-painting block, the hierarchical memory design and skip connections are among the most important contributions to performance. Based on the proposed block, all other components can be combined in a more effective way than SQUID. We attribute the improvements to better feature representation. With the memory design in this paper, the memory matrix is learned together with other model components, and learns a condensed feature representation for the whole training set. While with the original memory queue design in SQUID, the limited-size queue is only able to record a fraction of features in the training set, and these features are biased toward specific samples. Therefore, the improvement over SQUID comes from the memory features that better encode feature patterns in the training set.

Hyper-parameter robustness: The number of patch divisions, the topk value in Gumbel Shrinkage, and the number of memory patterns within a specific region of Memory Ma-

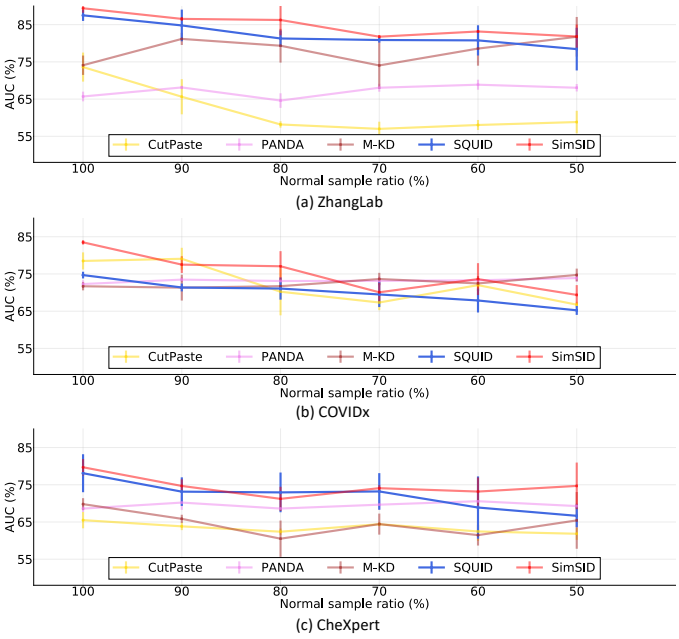


Fig. 9. Ablation study of mixing normal and abnormal samples in the training set. SimSID is robust to mixed training with different normal / abnormal ratios on the ZhangLab, COVIDx and CheXpert datasets.

trix are three important hyper-parameters of SimSID. Here, we conducted exhausted experiments on these parameters in Figure 8. Trials were first made on the number of patches from 1×1 to 8×8. When dividing input images into a single patch, space-aware settings are not triggered, hence yielding the worst performance. Although the spatial structures are relatively stable in most chest X-rays, certain deviations can still be observed. Therefore, with small patches, object parts in one patch can easily appear in adjacent patches and be misdetected as anomalies. Note that the best setting of patch number differs from SQUID, which attributes to the modification of the in-painting block. We found that Memory Matrix is less robust to spatial difference than Memory Queue. When segmenting an input image into non-overlapping patches, more patch segments lead to a smaller size for each patch, and, eventually, less inconsistency per patch. Therefore, the modified memory matrix in the in-painting block of SimSID benefits from more patches. The number of topk activations in Gumbel softmax also impacts the performances. By assembling the top-5 most similar patterns through Gumbel softmax, SimSID is able to achieve the best result. When replacing input features with the top-1 most similar pattern, SimSID suffers from a performance drop by -18.5% AUC. According to the AUC vs. number of patterns in each Memory Matrix region, we found that a small number of items is sufficient to support normal pattern querying in local regions and the best result is achieved by using merely 100 items per region. Degraded performance is observed at a greater number of items per region (>500).

5.3 Robustness to Abnormal Data in the Training Set

Strictly speaking, existing unsupervised anomaly detection methods (e.g., [13], [106]) are not unsupervised because they require a training set to be all "normal". To form this normal training set, image-level annotation as weak supervision is an implicit requirement. To the best of our knowledge, there is no work investigating the robustness of such "unsupervised" anomaly detection methods to the normal/abnormal ratio in real-world datasets. With disease-free sample ratio in the training set ranging from 100% to 50%, we have compared the robustness of SimSID with four competitive baselines (SQUID [83], CutPaste [97], PANDA [39] and M-KD [49]) that originally relies on a pure normal training set.

Figure 9 remarks that our proposed method is robust to the abnormal/normal training ratio up to 50% and remains AUC above 0.8 on the ZhangLab dataset by automatically omitting minority anatomical patterns. On the Stanford CheXpert and the COVIDx datasets, SimSID still achieves better or comparable AUC to abnormal training samples than the baseline models. We ask: *When the training set does not contain exclusively normal images, how does SimSID discriminate between abnormal and normal patches?* As described in §3, we divide an image into small patches and the model predicts every patch feature based on its eight surrounding patches. If a neighbor patch is abnormal, the other neighbors will contribute more to the in-painting process. Besides, abnormalities are often small, so the abnormal patches only account for a small proportion of an image, not to mention within the entire dataset. Since most cropped patches are normal, the abnormal patches would not have a serious effect, as evidenced by our robust detection results up to a 50% normal ratio. In contrast, CutPaste drops significantly as the percentage of disease-free images decreases; PANDA and M-KD can maintain their performance due to the use of pre-trained features. Interestingly, M-KD with mixed data even outperforms its vanilla training setting, although with considerable fluctuations. SQUID, on the other hand, benefits from the neighborhood in-painting design, but is still consistently worse than SimSID.

6 Conclusion

We present SimSID for unsupervised anomaly detection from radiography images. The assumption behind our design is that radiography imaging protocols focus on particular body regions, therefore producing images of great similarity and yielding recurrent anatomical structures across patients. SimSID exploits the structural consistency of chest anatomy with the help of *space-aware memory matrix* and *feature in-painting*. Qualitatively, we show that SimSID can taxonomize the ingrained anatomical structures into recurrent patterns; and in the inference, SimSID can identify anomalies (unseen/modified patterns) in the image. Quantitatively, SimSID surpasses the state of the arts in unsupervised anomaly detection by +8.0%, +5.0%, and +9.9% AUC scores on ZhangLab, COVIDx, and CheXpert benchmark datasets, respectively.

Acknowledgments

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award. We thank Xiaoxi Chen for annotating the Chest X-ray in Figure 1.

REFERENCES

[1] T. Zhao, K. Cao, J. Yao, I. Nogues, L. Lu, L. Huang, J. Xiao, Z. Yin, and L. Zhang, "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*, 2021, pp. 13 743–13 752.

[2] F. Haghighi, M. R. H. Taher, Z. Zhou, M. B. Gotway, and J. Liang, "Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning," *IEEE Transactions on Medical Imaging*, 2021.

[3] F. Haghighi, M. R. Hosseinzadeh Taher, Z. Zhou, M. B. Gotway, and J. Liang, "Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 137–147.

[4] Z. Zhou, M. B. Gotway, and J. Liang, "Interpreting medical images," in *Intelligent Systems in Medicine and Health*. Springer, 2022, pp. 343–371.

[5] Z. Zhou, "Towards annotation-efficient deep learning for computer-aided diagnosis," Ph.D. dissertation, Arizona State University, 2021.

[6] L. M. Smoger, C. K. Fitzpatrick, C. W. Clary, A. J. Cyr, L. P. Maletsky, P. J. Rullkoetter, and P. J. Laz, "Statistical modeling to characterize relationships between knee anatomy and kinematics," *Journal of Orthopaedic Research®*, vol. 33, no. 11, pp. 1620–1630, 2015.

[7] E. M. A. Anas, A. Rasoulian, A. Seitel, K. Darras, D. Wilson, P. S. John, D. Pichora, P. Mousavi, R. Rohling, and P. Abolmaesumi, "Automatic segmentation of wrist bones in ct using a statistical wrist shape + pose model," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1789–1801, 2016.

[8] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 737–745.

[9] Z. Zhou, J. Shin, R. Feng, R. T. Hurst, C. B. Kendall, and J. Liang, "Integrating active learning and transfer learning for carotid intima-media thickness video interpretation," *Journal of Digital Imaging*, vol. 32, no. 2, pp. 290–299, 2019.

[10] Y. Lu, W. Li, K. Zheng, Y. Wang, A. P. Harrison, C. Lin, S. Wang, J. Xiao, L. Lu, C.-F. Kuo *et al.*, "Learning to segment anatomical structures accurately from one exemplar," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 678–688.

[11] R. Feng, Z. Zhou, M. B. Gotway, and J. Liang, "Parts2whole: Self-supervised contrastive learning via reconstruction," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, 2020, pp. 85–95.

[12] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 384–393.

[13] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," *Medical Image Analysis*, vol. 69, p. 101952, 2021.

[14] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?" *Insights into imaging*, vol. 8, pp. 171–182, 2017.

[15] A. Zimek and E. Schubert, "Outlier detection," in *Encyclopedia of Database Systems*. Springer, 2017.

[16] S. Omar, A. Ngadi, and H. H. Jebur, "Machine learning techniques for anomaly detection: an overview," *International Journal of Computer Applications*, vol. 79, no. 2, 2013.

[17] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. C. Platt *et al.*, "Support vector method for novelty detection." in *NIPS*, vol. 12. Citeseer, 1999, pp. 582–588.

[18] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR 2011*. IEEE, 2011, pp. 3313–3320.

[19] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR 2011*. IEEE, 2011, pp. 3449–3456.

[20] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[21] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[22] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.

[23] D. Sidibe, S. Sankar, G. Lemaitre, M. Rastgoo, J. Massich, C. Y. Cheung, G. S. Tan, D. Milea, E. Lamoureux, T. Y. Wong *et al.*, "An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images," *Computer methods and programs in biomedicine*, vol. 139, pp. 109–117, 2017.

[24] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *International Conference on Learning Representations*, 2016.

[25] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *International Conference on Learning Representations*, 2017.

[26] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *International Conference on Learning Representations*, 2017.

[27] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[28] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[29] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *International Conference on Learning Representations*, 2018.

[30] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," *Medical Imaging with Deep Learning*, 2018.

[31] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *IEEE International Conference on Computer Vision*, 2019, pp. 191–200.

[32] Y. Tang, Y. Tang, Y. Zhu, J. Xiao, and R. M. Summers, "A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis," *Medical Image Analysis*, vol. 67, p. 101839, 2021.

[33] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical Image Analysis*, vol. 67, p. 101840, 2021.

[34] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339.

[35] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," *arXiv preprint arXiv:2111.09099*, 2021.

[36] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[37] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637.

[38] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, 2019.

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

[40] B. Nguyen, A. Feldman, S. Bethapudi, A. Jennings, and C. G. Willcocks, "Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1127–1131.

[41] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[42] M. Haselmann, D. P. Gruber, and P. Tabatabai, "Anomaly detection using deep learning based image completion," in *2018 17th*

IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018, pp. 1237–1242.

[43] J. Sato, Y. Suzuki, T. Wataya, D. Nishigaki, K. Kita, K. Yamagata, N. Tomiyama, and S. Kido, "Anatomy-aware self-supervised learning for anomaly detection in chest radiographs," *arXiv preprint arXiv:2205.04282*, 2022.

[44] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733–742.

[45] Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang, "Future frame prediction using convolutional vrnn for anomaly detection," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[46] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 588–13 597.

[47] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," *arXiv preprint arXiv:2111.08644*, 2021.

[48] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4183–4192.

[49] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 902–14 912.

[50] G. Pang, C. Aggarwal, C. Shen, and N. Sebe, "Editorial deep learning for anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2282–2286, 2022.

[51] Q. Hu, Y. Chen, J. Xiao, S. Sun, J. Chen, A. L. Yuille, and Z. Zhou, "Label-free liver tumor segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7422–7432.

[52] B. Li, Y.-C. Chou, S. Sun, H. Qiao, A. Yuille, and Z. Zhou, "Early detection and localization of pancreatic cancer by label-free tumor synthesis," *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023.

[53] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 152–21 164.

[54] Y. Zhang, X. Li, H. Chen, A. L. Yuille, Y. Liu, and Z. Zhou, "Continual learning for abdominal multi-organ and tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 35–45.

[55] M. A. Zuluaga, D. Hush, E. J. Delgado Leyton, M. H. Hoyos, and M. Orkisz, "Learning from only positive and unlabeled data to detect lesions in vascular ct images," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2011, pp. 9–16.

[56] M. A. Khan, T. Akram, Y.-D. Zhang, and M. Sharif, "Attributes based skin lesion detection and recognition: A mask rcnn and transfer learning-based deep learning framework," *Pattern Recognition Letters*, vol. 143, pp. 58–66, 2021.

[57] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[58] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[59] S. Zheng, J. Guo, X. Cui, R. N. Veldhuis, M. Oudkerk, and P. M. Van Ooijen, "Automatic pulmonary nodule detection in ct scans using convolutional neural networks based on maximum intensity projection," *IEEE transactions on medical imaging*, vol. 39, no. 3, pp. 797–805, 2019.

[60] N. U. Islam, S. Gehlot, Z. Zhou, M. B. Gotway, and J. Liang, "Seeking an optimal approach for computer-aided pulmonary embolism detection," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2021, pp. 692–702.

[61] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection–a survey," *arXiv preprint arXiv:2012.02364*, 2020.

[62] M. E. Tschuchnig and M. Gadermayr, "Anomaly detection in medical imaging–a mini review," *arXiv preprint arXiv:2108.11986*, 2021.

[63] M. Heer, J. Postels, X. Chen, E. Konukoglu, and S. Albarqouni, "The ood blind spot of unsupervised anomaly detection," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 286–300.

[64] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," *IEEE Winter Conference on Applications of Computer Vision*, 2022.

[65] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.

[66] S. Naval Marimont and G. Tarroni, "Implicit field learning for unsupervised anomaly detection in medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 189–198.

[67] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction," *BMC bioinformatics*, vol. 22, no. 2, pp. 1–20, 2021.

[68] H. Zhao, Y. Li, N. He, K. Ma, L. Fang, H. Li, and Y. Zheng, "Anomaly detection for medical images using self-supervised and translation-consistent features," *IEEE Transactions on Medical Imaging*, 2021.

[69] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*. PMLR, 2016, pp. 1378–1387.

[70] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1999–2007.

[71] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," *arXiv preprint arXiv:1703.03129*, 2017.

[72] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.

[73] S. Lee, J. Sung, Y. Yu, and G. Kim, "A memory network approach for story-based temporal summarization of 360 videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1410–1419.

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

[75] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 372–14 381.

[76] H. Lv, C. Chen, C. Zhen, C. Xu, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2021.

[77] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models for medical image analysis: A comprehensive survey," *arXiv preprint arXiv:2211.07804*, 2022.

[78] J. Linmans, G. Raya, J. van der Laak, and G. Litjens, "Diffusion models for out-of-distribution detection in digital pathology," *Medical Image Analysis*, p. 103088, 2024.

[79] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[80] S. Du, X. Wang, Y. Lu, Y. Zhou, S. Zhang, A. Yuille, K. Li, and Z. Zhou, "Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts," *arXiv preprint arXiv:2310.02906*, 2023.

[81] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2022, pp. 35–45.

[82] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, 2023.

[83] T. Xiang, Y. Zhang, Y. Lu, A. L. Yuille, C. Zhang, W. Cai, and Z. Zhou, "Squid: Deep feature in-painting for unsupervised anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 890–23 901.

[84] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[85] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[86] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[87] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Astrid, and S.-I. Lee, "An anomaly detection system via moving surveillance robots with human collaboration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2595–2601.

[88] D. Gong, Z. Zhang, J. Q. Shi, and A. van den Hengel, "Memory-augmented dynamic neural relational inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 843–11 852.

[89] K. Zhou, J. Li, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, J. Liu, and S. Gao, "Memorizing structure-texture correspondence for image anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[90] J. U. Kim, S. Park, and Y. M. Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3050–3059.

[91] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[92] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[93] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[94] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[96] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[97] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9664–9674.

[98] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 475–489.

[99] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, "Deep one-class classification via interpolated gaussian descriptor," *arXiv preprint arXiv:2101.10043*, 2021.

[100] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International conference on artificial neural networks*. Springer, 2011, pp. 52–59.

[101] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *2018 IEEE International conference on data mining (ICDM)*. IEEE, 2018, pp. 727–736.

[102] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2898–2906.

[103] Y. Liu, Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, and G. Carneiro, "Photoshopping colonoscopy video frames," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1–5.

[104] Y. Tian, G. Pang, F. Liu, Y. Chen, S. H. Shin, J. W. Verjans, R. Singh, and G. Carneiro, "Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 128–140.

[105] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," *arXiv preprint arXiv:2106.08265*, 2021.

[106] Y. Tian, G. Pang, Y. Liu, C. Wang, Y. Chen, F. Liu, R. Singh, J. W. Verjans, and G. Carneiro, "Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder," *arXiv preprint arXiv:2203.11725*, 2022.

[107] M. M. Rahman Siddiquee, T. Wu, and B. Li, "A2b-gan: Utilizing unannotated anomalous images for anomaly detection in medical image analysis," 2021.

[108] Y. Tian, F. Liu, G. Pang, Y. Chen, Y. Liu, J. Verjans, R. Singh, and G. Carneiro, "Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images," 2021.

[109] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[110] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[111] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[112] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, "Deep one-class classification via interpolated gaussian descriptor," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 383–392.

[113] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[114] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. 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*, 2017, pp. 2097–2106.

[115] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3544–3553.

[116] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1325–1334.

[117] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 249–258.

[118] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768.

[119] W. H. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, "Unsupervised brain imaging 3d anomaly detection and segmentation with transformers," *Medical Image Analysis*, p. 102475, 2022.