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

*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,zzhoud82}@jhu.edu; {yylu1989,alan.l.yuille}@gmail.com

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.
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 [30]–[34]. 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 [15], [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. nasal 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.
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 Inpainting 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 serves as a regularizer to prevent the student generator from collapsing (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].
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 $M$ is divided into blocks $\{M_{i,j} \in \mathbb{R}^{N \times C}\}$, each associated with a patch at location $(i, j)$, where $N$ and $C$ denote the number and the dimension of items, respectively. Let $z_{i,j} \in \mathbb{R}^{C}$ denote the feature of patch $(i, j)$, we obtain the augmented feature $\tilde{z}$ as follows:

$$\tilde{z}_{i,j} = \sum_{k=1}^{N} G(s^k)M_{i,j}^k,$$

where $s^k$ is the similarity score computed by dot product between the patch feature $z_{i,j}$ and the $k$-th memory item $M_{i,j}^k$. $G(\cdot)$ is the Gumbel-softmax operation, which shrinks the number of activated memory items. 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. In each generator layer, the feature map is up-sampled 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.
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 $E$, $G_t$, $G_s$, and $D$. An anomaly score ($A$) can be computed through:

$$A = \phi \left( \frac{D(G_s(E(I))) - \mu}{\sigma} \right),$$

where $\phi(\cdot)$ is the Sigmoid function, $\mu$ and $\sigma$ 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:

$$L_t = ||I - G_t(E(I))||^2, \quad L_s = ||I - G_s(E(I))||^2,$$

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

$$L_{dist} = \sum_{i=1}^{l} (z^t_i - z^s_i)^2,$$

where $l$ is the level of features used for knowledge distillation, $z_t$ and $z_s$ 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:

$$L_{gen} = \log(1 - D(G_s(E(I)))) + \log(D(I)).$$

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

$$L_{dis} = \log(D(I)) + \log(1 - D(G_s(E(I))))$$

In summary, our SimSID is trained to minimize the generative loss terms ($\lambda_t L_t + \lambda_s L_s + \lambda_{dist} L_{dist} + \lambda_{gen} L_{gen}$) and to maximize the discriminative loss term ($\lambda_{dis} L_{dis}$).

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 $128 \times 128$.

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 (valid set based on the official split). All images are resized to $128 \times 128$ 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.
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 [99]—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 \(\lambda = 0.01\), \(\lambda_1 = 10\), \(\lambda_{dist} = 0.001\), \(\lambda_{gen} = 0.005\), and \(\lambda_{disc} = 0.005\). We divided the input images in 4x4 non-overlapping patches generator. The architectures of our generators and discriminator are detailed in Figure 4.
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.

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-
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-
SimSID is robust to mixed training with different normal / abnormal ratios on the ZhangLab, COVIDx and CheXpert datasets. Here, we conducted exhausted experiments on these parameters in Figure 8. Trials were first made on the number of patches from $1 \times 1$ to $8 \times 8$. 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.


