SQUID: Deep Feature In-Painting for Unsupervised Anomaly Detection

Tiange Xiang[1] Yixiao Zhang[2] Yongyi Lu[2] Alan L. Yuille[2]
Chaoyi Zhang[1] Weidong Cai[1] Zongwei Zhou[2,*]

[1]University of Sydney [2]Johns Hopkins University

GitHub: https://github.com/tiangexiang/SQUID

Abstract

Radiography imaging protocols focus on particular body regions, therefore producing images of great similarity and yielding recurrent anatomical structures across patients. To exploit this structured information, we propose the use of Space-aware Memory Queues for In-painting and Detecting anomalies from radiography images (abbreviated as SQUID). We show that SQUID can taxonomize the ingrained anatomical structures into recurrent patterns; and in the inference, it can identify anomalies (unseen/modified patterns) in the image. SQUID surpasses 13 state-of-the-art methods in unsupervised anomaly detection by at least 5 points on two chest X-ray benchmark datasets measured by the Area Under the Curve (AUC). Additionally, we have created a new dataset (DigitAnatomy), which synthesizes the spatial correlation and consistent shape in chest anatomy. We hope DigitAnatomy can prompt the development, evaluation, and interpretability of anomaly detection methods.

1. Introduction

Vision tasks in photographic imaging and radiography imaging are different. For example, when identifying objects in photographic images, we assume translation invariance—a cat is a cat no matter if it appears on the left or right of the image. In radiography imaging, on the other hand, the relative location and orientation of a structure are important characteristics that allow the identification of normal anatomy and pathological conditions [20, 83]. Since radiography imaging protocols assess patients in a fairly consistent orientation, the generated images have great similarity across various patients, equipment manufacturers, and facility locations (see examples in Figure 1d). The consistent and recurrent anatomy facilitates the analysis of numerous critical problems and should be considered a significant advantage for radiography imaging [85]. Several investigations have demonstrated the value of this prior knowledge in enhancing Deep Nets' performance by adding location features, modifying objective functions, and constraining coordinates relative to landmarks in images [3, 47, 49, 69, 86]. Our work seeks to answer this critical question: *Can we exploit consistent anatomical patterns and their spatial information to strengthen Deep Nets' detection of anomalies from radiography images without manual annotation?*

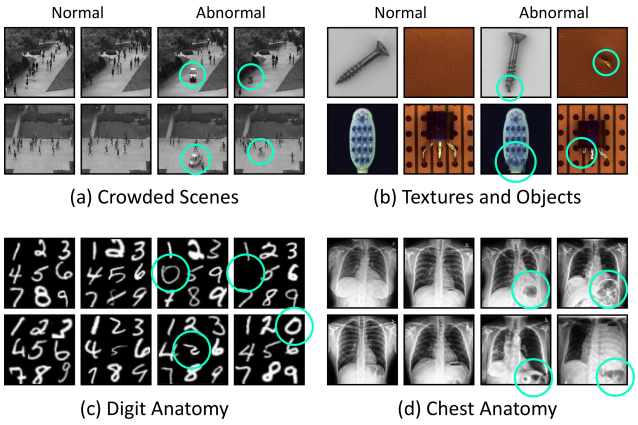


Figure 1. Anomaly detection in radiography images can be both easier and harder than in photographic images. It is easier because radiography images are spatially structured due to consistent imaging protocols. It is harder because anomalies in radiography images are subtle and require medical expertise to annotate.

Unsupervised anomaly detection only uses healthy images for model training and requires no other annotations such as disease diagnosis or localization [5]. As many as 80% of clinical errors occur when the radiologist misses the abnormality in the first place [7]. The 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. Unlike previous anomaly detection methods, we formulate the task as an in-painting task to exploit the anatomical consistency in appearance, position, and layout across radiography images. Specif-

*Corresponding author: Zongwei Zhou (zzhou82@jh.edu)

ically, we propose Space-aware Memory Queues for In-painting and Detecting anomalies from radiography images (abbreviated as SQUID). During training, our model can *dynamically* maintain a visual pattern dictionary by taxonomizing recurrent anatomical patterns based on their spatial locations. Due to the consistency in anatomy, the same body region across healthy images is expected to express similar visual patterns, which makes the total number of unique patterns manageable. During inference, since anomaly patterns do not exist in the dictionary, the generated radiography image is expected to be unrealistic if an anomaly is present. As a result, the model can identify the anomaly by discriminating the quality of the in-painting task. The success of anomaly detection has two basic assumptions [89]: *first*, anomalies only occur rarely in the data; *second*, anomalies differ from the normal patterns significantly.

We have conducted experiments on two large-scale, publicly available radiography imaging datasets. Our SQUID is significantly superior to predominant methods in unsupervised anomaly detection by over 5 points on the ZhangLab dataset [32]; remarkably, we have demonstrated a 10-point improvement over 13 recent unsupervised anomaly detection methods on the Stanford CheXpert dataset [29]. In addition, we have created a new dataset (DigitAnatomy) to elucidate *spatial correlation* and *consistent shape* of the chest anatomy in radiography (see Figure 1c). DigitAnatomy is dedicated to easing the development, evaluation, and interpretability of anomaly detection methods. The qualitative visualization clearly shows the superiority of our SQUID over the current state-of-the-art methods.

In summary, our contributions include: **(I)** the best performing unsupervised anomaly detection method for chest radiography imaging; **(II)** a synthetic dataset to promote anomaly detection research; **(III)** SQUID overcomes limitations in dominant unsupervised anomaly detection methods [1, 17, 35, 61, 82] by inventing Space-aware Memory Queue (§3.2), and Feature-level In-painting (§3.3).

2. Related Work

Anomaly detection in natural imaging. Anomaly detection is the task of identifying rare events that deviate from the distribution of normal data [52]. Early attempts include one-class SVM [64], dictionary learning [81], and sparse coding [11]. Due to the lack of sufficient samples of anomalies, later works typically formulate anomaly detection as an unsupervised learning problem [13, 26, 27, 37, 38, 42, 57, 67, 90]. 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 [9, 66, 71, 77, 87, 88] and identify anomalies by analyzing reconstruction errors. Density-based methods predict anomalies by estimating the normal data distribution (*e.g.* via VAEs [35] or GANs [2, 61, 62]). However, the learned distribution for normal images by these methods cannot explain the possible abnormalities. In this paper, we address these limitations by maintaining a visual pattern memory from homogeneous medical images. Several other previous works investigated the use of image inpainting 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 [22, 40, 51, 56, 79]. There are also plenty of works on detecting anomalies in video sequences [15, 45, 46]. Bergmann *et al.* [6] and Salehi *et al.* [59] proposed similar student-teacher networks, whereas our method utilizes such a structure to distillate input-aware features only, and the teacher network is completely disabled during inference.

Anomaly detection in medical imaging. Anomaly detection in the medical domain is usually approached on a per-pathology basis. Supervised learning based methods [5, 44, 63] are commonly adopted to detect specific types of abnormalities, such as lesions [91], pathologies [33], tumors [4], and nodules [84]. Recent unsupervised methods have been proposed to detect anomalies in general [5, 25, 66], With the help of GANs, anomaly detection can be achieved with *weak* annotations. In AnoGAN [62], the discriminator was heavily over-fitted to the normal image distribution to detect the anomaly. Subsequently, f-AnoGAN [61] was proposed to improve computational efficiency. Marimont *et al.* [50] designed an auto-decoder 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.* [21] 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 [82] 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. Differing from photographic images, radiography imaging protocols produce images with consistent anatomical patterns, which are much more challenging to detect due to subtle imaging clues and overlapping anatomic structures (Figure 1). Unlike most existing works, we present a novel method that explicitly harnesses the radiography images' properties, dramatically improving the performance in anomaly detection from radiography images.

Memory networks. Incorporating memory modules into neural networks has been demonstrated to be effective for many tasks [8, 16, 31, 36, 39]. Adopting Memory Matrix

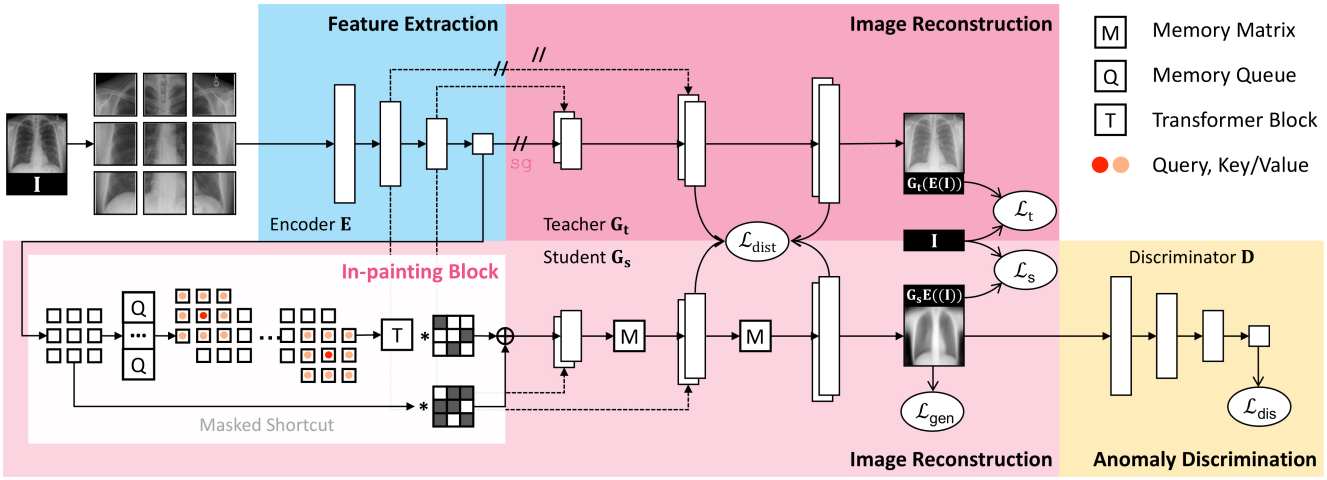


Figure 2. **SQUID.** We divide an input image into $N \times N$ non-overlapping patches and feed them into the encoder for feature extraction. Two generators will be trained to reconstruct the original image. Along with the reconstruction, a dictionary of anatomical patterns will be created and updated dynamically via a novel Memory Queue (§3.2); The teacher generator directly uses the features extracted by the encoder; the student generator uses the features augmented by our 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, it can also be used to detect anomalies in test images (§3.4).

for unsupervised anomaly detection was first proposed in MemAE [17]. In addition to auto-encoding (AE), an extra Memory Matrix was introduced 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*. [53] introduced a non-learnable memory module that can be updated with inputs. Note that although our proposed Memory Queue also does not require any gradients, our method differs significantly in its usage purpose and updating rules. Considering the extra memory usage in existing methods, Lv *et al*. [48] proposed a dynamic prototype unit that encodes normal dynamics on the fly, while consuming little additional memory. In this paper, we overcome the limitations of the Memory Matrix and propose an effective yet efficient Memory Queue for unsupervised anomaly detection in radiography images.

3. SQUID

3.1. Overview

(1) *Feature extraction.* We divide the input image into $N \times N$ non-overlapping patches and feed them into an encoder for feature extraction. The extracted features will be used for image reconstruction. Practically, the encoder can be any backbone architectures [14, 70]; we adopt basic Convolutions and Pooling layers in this work for simplicity.

(2) *Image reconstruction.* We introduce teacher and student generators to reconstruct the original image. Along with the reconstruction, a dictionary of anatomical patterns will be created and updated dynamically as a **Memory Queue** (§3.2). Specifically, the teacher generator directly reconstructs the image using the features extracted by the encoder (essentially an auto-encoder [58]). The student generator, on the other hand, using the features augmented by our **in-painting block** (§3.3). The teacher and student generators are coupled through a knowledge distillation paradigm [28] at all of the up-sampling levels. The objective of the student generator is to reconstruct a normal image from the augmented features, which will then be used for anomaly discrimination (§3.4); while the teacher generator[1] serves as a regularizer that prevents the student from constantly generating the same normal image.

(3) *Anomaly discrimination.* Following the adversarial learning [61, 62], we employ a discriminator to assess whether the generated image is real or fake. Only the student generator will receive the gradient derived from the discriminator. The two generators and the discriminator are competing against each other until they converge to an equilibrium. Once trained, the discriminator can be used to detect anomalies in test images (§3.4).

3.2. Inventing Memory Queue as Dictionary

Motivation. The Memory Matrix was introduced by Gong *et al*. [17] and has since been widely adopted in unsupervised anomaly detection [18, 45, 78]. To forge a "normal" appearance, features are *augmented* by weighted averaging

[1] We disabled the backpropagation between the teacher and encoder by stop-gradient [23] and showed its empirical benefit in Table 2.

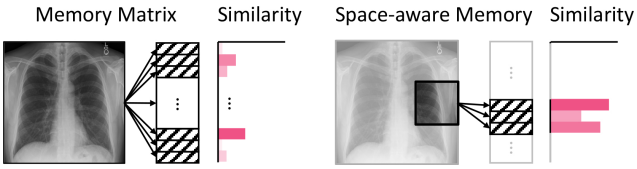


Figure 3. **Space-aware memory.** For unique encoding of location information, we restrict each patch to be only accessible by a non-overlapping region in the memory.

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

Space-aware memory. To harness the spatial information, we pass the divided small patches, instead of the whole image, into the model. These patches are associated with unique location identifiers of the original image. We seek to build the relationship between the patch location and memory region, by restricting the search space in Memory Matrix to the patch-corresponded non-overlapping segments only. That is, a patch at a particular location can only access a corresponding segment in the whole Memory Matrix (illustrated in Figure 3). We refer to this new strategy as "space-aware memory" because it enables explicit encoding of the spatial information into Memory Matrix. Space-aware memory can also accelerate the augmentation speed compared with [17] as it no longer goes through the entire Memory Matrix to assemble similar features.

Memory queue. In learning-based Memory Matrix [17], "normal patterns" are forged by combining learned basis in the matrix. However, there is always a distribution discrepancy between the basic combinations and the actual image features. This disparity makes it hard for the subsequent image generation. To address this issue, we propose a Memory Queue to store *real* image features during model training, therefore presenting an identical distribution to the image features. Specifically, it directly copies previously seen features into a queue structure during training[2]. Once trained, Memory Queue can be used as a *dictionary* of normal anatomical patterns. In Figure 4, we show t-SNE visualizations to validate that the learned basis in Memory Matrix (blue dots) distributes differently from the actual image features of the training set (gray dots). In contrast, the stored image features in our Memory Queue (red dots) are

[2]In practice, copying features into the queue at every training iteration demands considerable computational time. Supposing N patterns in the queue and M training iterations, the sampling strategy in [17] demands an $\mathcal{O}(NM)$ time complexity. We implement it more efficiently: at each iteration, the current batch of features will be copied into the queue for *only once* (Figure 5c), yielding a linear complexity of $\mathcal{O}(M + cM)$ with the copy-and-paste operation in a constant time c. We follow the first-in-first-out (FIFO) paradigm to update the queue continuously.

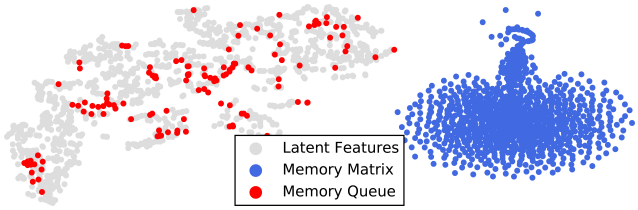


Figure 4. t-SNE visualizations of patterns in Memory Matrix, Memory Queue, and patch features of the training samples [73]. Patterns in the Memory Matrix are far away from the distribution of patch features, while patterns in the Memory Queue (as copies of previously seen features) share a similar distribution.

in an identical distribution to the actual ones.

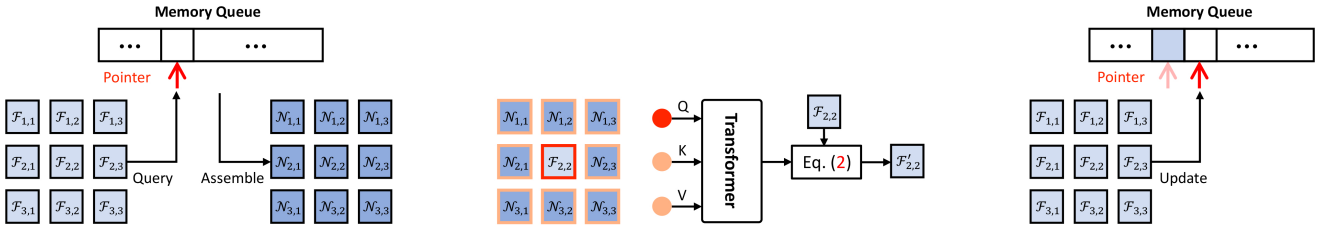
Gumbel shrinkage. Controlling the number of activated patterns in the memory has proven to be advantageous for anomaly detection [17, 19]. However, setting a hard shrinkage threshold fails to adapt to cases where no suitable entries can be found in the memory. One natural workaround is to activate the top-k similar patterns in the memory. However, this strategy restricts the gradient flow to only the top-k memory entries, while the rest inactivated ones could not receive any gradients and be updated as expected. To extend gradients to all patterns in the memory, inspired by Jang [30], we present a *Gumbel Shrinkage* schema:

$$\mathbf{w}' = \texttt{sg}(\texttt{hs}(\mathbf{w}, \texttt{topk}(\mathbf{w})) - \phi(\mathbf{w})) + \phi(\mathbf{w}), \quad (1)$$

where $\mathbf{w}$ denotes the similarity between the image features and entries in Memory, $\texttt{sg}(\cdot)$ the stop-gradient operation, $\texttt{hs}(\cdot, t)$ the hard shrinkage operator with threshold t, and $\phi(\cdot)$ the Softmax function. In the forward pass, Gumbel Shrinkage ensures the combination of the top-k most similar entries in the memory; During the back-propagation, Gumbel Shrinkage essentially functions as Softmax. We apply Gumbel Shrinkage to both Memory Queue and Memory Matrix in our framework.

3.3. Formulating Anomaly Detection as In-painting

Motivation. Image in-painting [41, 54] was initially proposed to recover corrupted image regions with neighboring context. Following the above intuition, we propose to achieve anomaly detection via in-painting anomalous radiography patterns into healthy ones. When in-painting pixels in the image space, recovered regions have been usually seen to associate with boundary artifacts, particularly when using Deep Nets [43]. These undesired artifacts are responsible for numerous false positives when formulating anomaly detection as a pixel-level in-painting task [66, 87]. To alleviate this issue, we achieve the in-painting task at the feature level instead. Latent features are better invariant to pixel-level noise, rotation, and translation, therefore are more suitable for subsequent anomaly detection.



(a) Augment $\mathcal{F}$ to $\mathcal{N}$ (b) Aggregate $\mathcal{F}$ and $\mathcal{N}$ (c) Copy $\mathcal{F}$ to Memory Queue

Figure 5. **Three-step workflow of our in-painting block.** (a) Each non-overlapping patch feature $\mathcal{F}$ is queried to a unique region in Memory Queue, and the most similar items are assembled to $\mathcal{N}$. (b) Each center patch feature $\mathcal{F}$ and its eight neighbors $\mathcal{N}$ are used as query and key/value, respectively, to a Transformer layer for in-painting. (c) Each Memory Queue region copies its corresponding patch features $\mathcal{F}$ into the memory by maintaining a pointer. Note that this step is only performed during training.

In-painting block. We integrate our Memory Queue inside a novel in-painting block to perform feature-space in-painting. The block starts with a Memory Queue that augments $w \times h$ non-overlapping patch features $\mathcal{F}_{\{(1,1),\cdots,(w,h)\}}$ into their most similar "normal" patterns $\mathcal{N}_{\{(1,1),\cdots,(w,h)\}}$ (Figure 5a). Since $\mathcal{N}$ is assembled by features extracted from previous training data, $\mathcal{N}$ is not subject to the current input image. To recap the characteristics of the input image, we aggregate both patch features $\mathcal{F}$ and their augmented features $\mathcal{N}$ using a transformer block [74]. In details, for each patch $\mathcal{F}_{i,j}$, its spatially adjacent eight augmented ones $\mathcal{N}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}}$ are used as conditions to refine $\mathcal{F}_{i,j}$ (Figure 5b). The query token of the transformer block is flattened $\mathcal{F}_{(i,j)} \in \mathcal{R}^{1\times *}$ and key/value tokens are $\mathcal{N}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}} \in \mathcal{R}^{8\times *}$. At the start and the end of our in-painting block, we apply an extra pair of point-wise convolutions (1×1 convolutional kernel) [24].

Masked shortcut. We employ a shortcut within the in-painting block to better aggregate features and ease optimization. Our empirical study shows that a direct residual connection downgrades the effectiveness of the in-painting block (Appendix B). Inspired by Xiang *et al*. [76], we utilize a random binary mask to gate shortcut features during training (Figure 5b). As such, given the input patch features $\mathcal{F}$, the output of the in-painting block is obtained by:

$$\mathcal{F}' = (1-\delta)\cdot\mathcal{F} + \delta\cdot\texttt{inpaint}(\mathcal{F}), \qquad (2)$$

where $\texttt{inpaint}(\cdot)$ is the designed in-painting block, $\delta \sim$ *Bernoulli*(ρ) is a binary variable with ρ the gating probability. After obtaining $\mathcal{F}'$ at each training step, the originally $\mathcal{F}$ are then copied to update the memory (Figure 5c). During inference, we disable the shortcut completely such that $\mathcal{F}' = \texttt{inpaint}(\mathcal{F})$ for deterministic predictions.

3.4. Anomaly Discrimination

Our discriminator can detect anomalies by assessing the quality of the reconstructions—normal if realistic; abnormal otherwise. It is because the generator was trained on normal images, so Memory Queues only store normal patterns. During inference, since abnormal patterns were never present in Memory Queues, the reconstructed image is expected to appear unrealistic.

Our in-painting block focuses on augmenting any patch feature (either normal or abnormal) into similar "normal" features. The student generator then reconstructs a "normal" image based on the "normal" features. The teacher generator is used to prevent the student from generating the same image regardless of inputs. Once trained, the semantic (rather than pixel-level) difference between the input and the student generator's reconstructed image is expected to be small if normal and big otherwise. We, therefore, delegate the optimized discriminator network for alerting anomalies perceptually. 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: $A = \phi(\frac{\mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I})))-\mu}{\sigma})$, where $\phi(\cdot)$ is the Sigmoid function, μ and σ are the mean and standard deviation of anomaly scores calculated on training samples.

3.5. Loss Function

SQUID 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, $\mathcal{L}_\text{t} = (\mathbf{I} - \mathbf{G}_\text{t}(\mathbf{E}(\mathbf{I})))^2$ and $\mathcal{L}_\text{s} = (\mathbf{I} - \mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I})))^2$ for the teacher and student generators, respectively, where $\mathbf{I}$ denotes the input image. Following the knowledge distillation paradigm, we apply a distance constraint between the teacher and student generators to all levels of features: $\mathcal{L}_\text{dist} = \sum_{i=1}^{l}(\mathcal{F}_\text{t}^i - \mathcal{F}_\text{s}^i)^2$, where l is the level of features used for knowledge distillation, $\mathcal{F}_\text{t}$ and $\mathcal{F}_\text{s}$ are the intermediate features in the teacher and student generators, respectively. In addition, we employ an adversarial loss (similar to DCGAN [55]) to improve the quality of the image generated by the student generator. Specifically, the following equation is minimized: $\mathcal{L}_\text{gen} = \log(1-\mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I}))))$. The discriminator seeks to maximize the average of the proba-

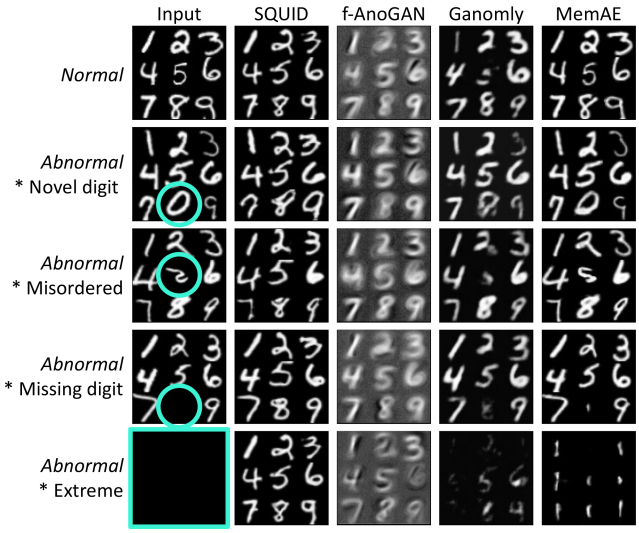


Figure 6. **Reconstruction results on DigitAnatomy.** Our feature-level in-painting is more robust to amplified noise and pixel variance than the existing pixel-level in-painting methods. More visualization can be found in Appendix D.

bility for real images and the inverted probability for fake images: $\mathcal{L}_{\text{dis}} = \log(\mathbf{D}(\mathbf{I})) + \log(1 - \mathbf{D}(\mathbf{G}_{\text{s}}(\mathbf{E}(\mathbf{I}))))$. In summary, SQUID 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. New Benchmark

DigitAnatomy. We have created a synthetic dataset to verify our main idea, wherein the human anatomy is translated into Arabic digits one to nine in an in-grid placement (see examples in Figure 1 and Figure 6). The images containing digits in the correct order are considered "normal"; otherwise, they are considered "abnormal". The types of simulated abnormalities include missing, misordered, flipped, and zero digit(s). DigitAnatomy is particularly advantageous for radiography imaging for three reasons. *First*, it simulates two unique properties of radiography images, *i.e.* spatial correlation and consistent shape. *Second*, annotating radiography images demands specialized expertise, but digits are easier for problem debugging. *Third*, the ground truth of the simulated anomaly is readily accessible in DigitAnatomy, whereas it is hard to collect sufficient examples for each abnormal type in radiography images. The pseudocode for creating DigitAnatomy is in Appendix C.

4.2. Public Benchmarks

ZhangLab Chest X-ray [32]. This dataset contains healthy and pneumonia (as anomaly) images, *officially* split into training and testing sets. The training set consists of 1,349 normal and 3,883 abnormal images; the testing 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 hyper-parameter tuning. We resized all the images to 128×128.

Table 1. Benchmark results on the test sets of the two datasets.

ZhangLab	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Auto-Encoder	-	59.9	63.4	77.2
VAE [35]	Arxiv'13	61.8	64.0	77.4
Ganomaly [1]	ACCV'18	78.0	70.0	79.0
f-AnoGAN [61]	MIA'19	75.5	74.0	81.0
MemAE [17]	ICCV'19	77.8±1.4	56.5±1.1	82.6±0.9
MNAD [53]	CVPR'20	77.3±0.9	73.6±0.7	79.3±1.1
SALAD [82]	TMI'21	82.7±0.8	75.9±0.9	82.1±0.3
CutPaste [40]	CVPR'21	73.6±3.9	64.0±6.5	72.3±8.9
PANDA [56]	CVPR'21	65.7±1.3	65.4±1.9	66.3±1.2
M-KD [59]	CVPR'21	74.1±2.6	69.1±0.2	62.3±8.4
IF 2D [50]	MICCAI'21	81.0±2.8	76.4±0.2	82.2±2.7
PaDiM [12]	ICPR'21	71.4±3.4	72.9±2.4	80.7±1.2
IGD [10]	AAAI'22	73.4±1.9	74.0±2.2	80.9±1.3
SQUID	-	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**
CheXpert	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Ganomaly [1]	ACCV'18	68.9±1.4	65.7±0.2	65.1±1.9
f-AnoGAN [61]	MIA'19	65.8±3.3	63.7±1.8	59.4±3.8
MemAE [17]	ICCV'19	54.3±4.0	55.6±1.4	53.3±7.0
CutPaste [40]	CVPR'21	65.5±2.2	62.7±2.0	60.3±4.6
PANDA [56]	CVPR'21	68.6±0.9	66.4±2.8	65.3±1.5
M-KD [59]	CVPR'21	69.8±1.6	66.0±2.5	63.6±5.7
SQUID	-	**78.1±5.1**	**71.9±3.8**	**75.9±5.7**

Stanford CheXpert [29]. 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 images, 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 were used for testing. We used the same hyper-parameters found in the ZhangLab experiments.

4.3. Baselines and Metrics

We considered a total number of **13** major baselines for direct comparison: Auto-Encoder, VAE [35]—the classic UAD methods; Ganomaly [1], f-AnoGAN [61], IF [50], SALAD [82]—the current state of the arts for medical imaging; and MemAE [17], CutPaste [40], M-KD [60], PANDA [56], PaDiM [12], IGD [10]—the most recent UAD methods. We evaluated performance using standard metrics: Receiver Operating Characteristic (ROC) curve, Area Under the ROC Curve (AUC), Accuracy (Acc), and F1-score (F1). Unless explicitly specified, we trained all models from scratch for at least *three* times independently.

4.4. Implementation Details

We utilized common data augmentation strategies such as random translation within the $[-0.05, +0.05]$ range and a random scaling of $[0.95, 1.05]$. The Adam [34] 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

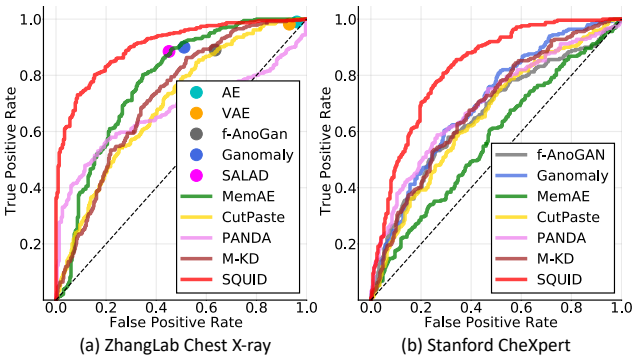


Figure 7. ROC curves comparison on the two datasets.

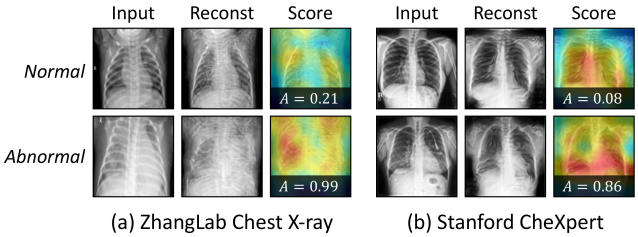


Figure 8. Reconstruction results of SQUID on the two datasets, associated with the corresponding anomaly scores (defined in §3.4). A larger score indicates a higher probability of being abnormal. More visualization can be found in Appendix D.

generator and the discriminator and then decayed to $2e$-5 in 1000 epochs following the cosine annealing scheduler. The discriminator is trained at every iteration, while the generator is trained every two iterations. We set 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 divide the input images in 2×2 non-overlapping patches, fix the shortcut mask probability at $\rho = 95\%$, and activate only the top 5 similar patterns in the Gumbel Shrinkage. The impact of these hyper-parameters is studied in §5.3. The architectures of our generators and discriminator are detailed in Appendix A.

5. Results

5.1. Interpreting SQUID on DigitAnatomy

Figure 6 presents qualitative results on DigitAnatomy to examine the capability of image reconstruction and to interpret the mistakes made by existing methods [1, 17, 61]. We deliberately inject anomalies (*e.g.* novel, misordered, missing digits) into normal images (highlighted in light blue) and test if the model can reconstruct their normal counterparts. To raise the task difficulty, we also assess the reconstruction quality from a blank image (as an extreme case). In general, the images reconstructed by our SQUID carry more meaningful and indicative information than other baseline methods. It is mainly attributed to our *space-aware* memory, with which the resulting dictionary is associated with unique patterns as well as their spatial information. Once an anomaly arises (*e.g.* missing digit), the in-painting block will augment the abnormal feature to its normal counterpart by assembling top-k most similar patterns from the dictionary. Other methods, however, do not possess this ability, so they reconstruct defective images. For instance, GAN-based methods (f-AnoGAN and Ganomly) tend to reconstruct an exemplar image averaged from the training examples. MemAE performs relatively better due to its Memory Matrix, but it does not work well for the anomaly of missing digits and completely fails on the extreme anomaly attack.

5.2. Benchmarking SQUID on Chest Radiography

Our SQUID was mainly evaluated on two large-scale benchmarks: ZhangLab Chest X-ray and Stanford CheXpert and compared with a wide range of state-of-the-art counterparts. According to Table 1, SQUID achieves the most promising result on all metrics for both datasets. Specifically, SQUID outperforms the runner-up counterparts by at least 5% in AUC, 5% in Accuracy. The highest F1 scores SQUID achieved, along with the ROC curves shown in Figure 7, demonstrate that our method yields the best trade-off between sensitivity and specificity. Overall, the significant improvements observed with SQUID proved the effectiveness of our proposed techniques in this work. In Figure 8, we visualize the reconstruction results of SQUID on exemplary normal and abnormal images in the two datasets. For normal cases, SQUID can easily find a similar match in Memory Queue, achieving 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 [65]) to indicate the most likely regions to appear anomalous. As a result, the reconstructed healthy images yield much lower anomaly scores than the diseased ones, validating the effectiveness of SQUID.

Limitation. We found SQUID in its current form, is not able to *localize* anomalies at the pixel level precisely. It is understandable because our SQUID is an unsupervised method, requiring zero manual annotation for normal/abnormal images, unlike [60, 68, 72, 75, 80]. Those methods that compute pixel-level residuals for anomaly detection suffer from amplified noise in the input and reconstructed output. Our in-painting strategy, however, is performed at the feature level and is more robust to pixel-level variance.

5.3. Ablating Key Properties in SQUID

Component study. We examine the impact of components in SQUID by taking each one of them out of the entire

Table 2. Performance benefits from all the components in SQUID.

Method	AUC (%)	Acc (%)	F1 (%)
w/o Space-aware Memory	77.6±0.5	75.5±0.5	82.5±0.6
w/o In-painting Block	80.9±2.1	75.8±1.5	81.6±1.3
w/o Gumbel Shrinkage	81.1±0.9	77.6±0.9	81.3±0.8
w/o Knowledge Distillation	81.2±0.8	75.2±0.7	81.3±0.8
w/o Stop Gradient	81.7±4.3	76.7±2.8	82.5±1.6
w/o Memory Queue	82.5±1.1	78.6±0.9	81.7±1.1
w/o Masked Shortcuts	82.5±1.3	76.4±0.8	82.3±1.1
w/o Decoder Memory	82.9±1.2	77.4±1.1	81.2±0.5
Full SQUID	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**

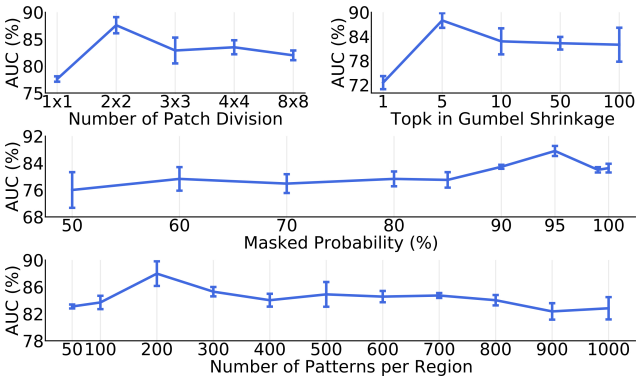


Figure 9. Hyper-parameter ablations.

framework. Table 2 shows that each component accounts for at least 5% performance gain. The space-aware memory (+10.0%) and in-painting block (+6.7%) are the top 2 most significant contributors, which underline our motivation and justification of the method development (§3.2 and §3.3). Although replacing Memory Queue with Memory Matrix could maintain a decent result (only dropped 5.1%), our Memory Queue presents a more trustworthy recovery of "normal" patterns in the image than Memory Matrix (MemAE [17]), evidenced by Figure 6.

Hyper-parameter robustness. After selecting the best hyper-parameters on the validation set, we here report the inference results on the testing set to study the robustness of different hyper-parameters in Figure 9. When input images are divided into a single patch, space-aware settings are not triggered, therefore yielding the worst performance. Although the spatial structures are relatively stable in most chest radiography, 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. The number of topK activations in Gumbel softmax also impacts the performances. According to the AUC vs. the number of patterns in each Memory Queue region, we found that a small number of items is sufficient to support normal pattern querying in local regions. The best result is achieved by merely 200 items per region. When the item number exceeds 500 per region, AUC scores begin to drop continuously. AUC vs. the mask probability ρ was

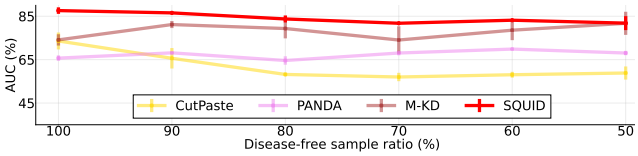


Figure 10. Results of mixing normal/abnormal training samples.

further plotted to verify that enabling a limited number of feature skips ($\rho = 95\%$) yields the best AUC score. The effectiveness of the in-painting will severely deteriorate if more features are allowed to be skipped ($\rho < 90\%$).

Disease-free training requirement? Unsupervised methods for medical anomaly detection are uncommon because the so-called UAD methods are *not* "unsupervised"—they must be trained on disease-free images only (*e.g.* [5]). In practice, cleaning up disease-free images relies on manual annotation (essentially, image-level healthy/diseased labels). With disease-free sample ratio in the training set ranging from 100% to 50%, we have compared the robustness of SQUID with three competitive baselines (CutPaste [40], PANDA [56] and M-KD [60]). Figure 10 remarks that our proposed memory queue can tolerate the disease/healthy training ratio up to 50% by automatically omitting minority anatomical patterns. In contrast, CutPaste drops significantly as the percentage of normal images decreases; PANDA and M-KD can maintain the 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.

6. Conclusion

We present SQUID for unsupervised anomaly detection from radiography images. Qualitatively, we show that SQUID can taxonomize the ingrained anatomical structures into recurrent patterns; and in the inference, SQUID can identify anomalies accurately. Quantitatively, SQUID is superior to predominant methods by over 5 points AUC on the ZhangLab dataset and 10 points AUC on the Stanford CheXpert dataset. The outstanding results are attributable to our observation: *Radiography imaging protocols focus on particular body regions, therefore producing images of great similarity and yielding recurrent anatomical structures across patients.* We synthesized the DigitAnatomy dataset to resemble key attributes of chest anatomy in radiography images for prompting future method development.

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and partially by the Patrick J. McGovern Foundation Award. We appreciate the constructive suggestions from Yingda Xia, Yixiao Zhang, Jessica Han, Yingwei Li, Bowen Li, Huiyu Wang, Adam Kortylewski, and Sonomi Oyagi.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 2, 6, 7, 14, 15

[2] Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *Image Processing Medical Imaging*, volume 10133, page 101330G. International Society for Optics and Photonics, 2017. 2

[3] Emran Mohammad Abu Anas, Abtin Rasoulian, Alexander Seitel, Kathryn Darras, David Wilson, Paul St John, David Pichora, Parvin Mousavi, Robert Rohling, and Purang Abolmaesumi. Automatic segmentation of wrist bones in ct using a statistical wrist shape + pose model. *IEEE transactions on medical imaging*, 35(8):1789–1801, 2016. 1

[4] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin 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. 2

[5] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021. 1, 2, 8

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 2

[7] Adrian P Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, 8:171–182, 2017. 1

[8] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4080–4088, 2018. 2

[9] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *Medical Imaging with Deep Learning*, 2018. 2

[10] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. *arXiv preprint arXiv:2101.10043*, 2021. 6

[11] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011. 2

[12] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 6

[13] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020. 3

[15] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2

[16] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2

[17] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den 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*, pages 1705–1714, 2019. 2, 3, 4, 6, 7, 8, 13, 14, 15

[18] Dong Gong, Zhen Zhang, Javen Qinfeng Shi, and Anton van den Hengel. Memory-augmented dynamic neural relational inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11843–11852, 2021. 3

[19] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 4

[20] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 2021. 1

[21] Changhee Han, Leonardo Rundo, Kohei Murao, Tomoyuki Noguchi, Yuki Shimahara, Zoltán Ádám Milacski, Saori Koshino, Evis Sala, Hideki Nakayama, and Shin'ichi Satoh. Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *BMC bioinformatics*, 22(2):1–20, 2021. 2

[22] Matthias Haselmann, Dieter P Gruber, and Paul Tabatabai. Anomaly detection using deep learning based image completion. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1237–1242. IEEE, 2018. 2

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[25] Matthäus Heer, Janis Postels, Xiaoran Chen, Ender Konukoglu, and Shadi Albarqouni. The ood blind spot of unsupervised anomaly detection. In *Medical Imaging with Deep Learning*, pages 286–300. PMLR, 2021. 2

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2016. 2

[27] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2018. 2

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[29] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 2, 6

[30] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[31] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017. 2

[32] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 2, 6

[33] Muhammad Attique Khan, Tallha Akram, Yu-Dong Zhang, and Muhammad Sharif. Attributes based skin lesion detection and recognition: A mask rcnn and transfer learning-based deep learning framework. *Pattern Recognition Letters*, 143:58–66, 2021. 2

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 6

[36] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387. PMLR, 2016. 2

[37] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations*, 2017. 2

[38] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2

[39] Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee 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*, pages 1410–1419, 2018. 2

[40] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 2, 6, 8

[41] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 4, 13

[42] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations*, 2017. 2

[43] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 4

[44] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*, 2023. 2

[45] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing 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*, pages 13588–13597, 2021. 2, 3

[46] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

[47] Yuhang Lu, Weijian Li, Kang Zheng, Yirui Wang, Adam P Harrison, Chihung Lin, Song Wang, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Learning to segment anatomical structures accurately from one exemplar. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 678–688. Springer, 2020. 1

[48] Hui Lv, Chen Chen, Cui Zhen, Chunyan Xu, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2021. 3

[49] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018. 1

[50] Sergio Naval Marimont and Giacomo Tarroni. Implicit field learning for unsupervised anomaly detection in medical im-

ages. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–198. Springer, 2021. 2, 6

[51] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1127–1131. IEEE, 2021. 2

[52] Salima Omar, Asri Ngadi, and Hamid H Jebur. Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, 79(2), 2013. 2

[53] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 3, 6

[54] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 4, 13

[55] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5

[56] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. 2, 6, 8

[57] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 2

[58] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 3

[59] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad Hossein Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. 2020. 2, 6

[60] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14902–14912, 2021. 6, 7, 8

[61] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 2019. 2, 3, 6, 7, 14, 15

[62] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2, 3

[63] Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015. 2

[64] Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588. Citeseer, 1999. 2

[65] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[66] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019. 2, 4

[67] Desire Sidibe, Shrinivasan Sankar, Guillaume Lemaitre, Mojdeh Rastgoo, Joan Massich, Carol Y Cheung, Gavin SW Tan, Dan Milea, Ecosse Lamoureux, Tien 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*, 139:109–117, 2017. 2

[68] Krishna Kumar Singh and Yong Jae 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)*, pages 3544–3553. IEEE, 2017. 7

[69] Lowell M Smoger, Clare K Fitzpatrick, Chadd W Clary, Adam J Cyr, Lorin P Maletsky, Paul J Rullkoetter, and Peter J Laz. Statistical modeling to characterize relationships between knee anatomy and kinematics. *Journal of Orthopaedic Research®*, 33(11):1620–1630, 2015. 1

[70] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3

[71] Youbao Tang, Yuxing Tang, Yingying Zhu, Jing Xiao, and Ronald M Summers. A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis. *Medical Image Analysis*, 67:101839, 2021. 2

[72] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald 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*, pages 249–258. Springer, 2018. 7

[73] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 5

[75] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[76] Tiange Xiang, Chaoyi Zhang, Yang Song, Siqi Liu, Hongliang Yuan, and Weidong Cai. Partial graph reasoning for neural network regularization. *arXiv preprint arXiv:2106.01805*, 2021. 5

[77] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. *IEEE Winter Conference on Applications of Computer Vision*, 2022. 2

[78] Muhammad Zaigham Zaheer, Arif Mahmood, M Haris Khan, Marcella Astrid, and Seung-Ik Lee. An anomaly detection system via moving surveillance robots with human collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2595–2601, 2021. 3

[79] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2

[80] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 7

[81] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011. 2

[82] He Zhao, Yuexiang Li, Nanjun He, Kai Ma, Leyuan Fang, Huiqi Li, and Yefeng Zheng. Anomaly detection for medical images using self-supervised and translation-consistent features. *IEEE Transactions on Medical Imaging*, 2021. 2, 6

[83] Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling 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*, pages 13743–13752, 2021. 1

[84] Sunyi Zheng, Jiapan Guo, Xiaonan Cui, Raymond NJ Veldhuis, Matthijs Oudkerk, and Peter MA Van Ooijen. Automatic pulmonary nodule detection in ct scans using convolutional neural networks based on maximum intensity projection. *IEEE transactions on medical imaging*, 39(3):797–805, 2019. 2

[85] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021. 1

[86] Zongwei Zhou, Jae Shin, Ruibin Feng, R Todd Hurst, Christopher B Kendall, and Jianming Liang. Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. *Journal of digital imaging*, 32(2):290–299, 2019. 1

[87] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 2, 4, 13

[88] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 2

[89] Arthur Zimek and Erich Schubert. Outlier detection. In *Encyclopedia of Database Systems*. Springer, 2017. 2

[90] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2

[91] Maria A Zuluaga, Don Hush, Edgar JF Delgado Leyton, Marcela Hernández Hoyos, and Maciej 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*, pages 9–16. Springer, 2011. 2

A. Architectures of SQUID

Our SQUID consists of an encoder, a student (main) generator, a teacher generator, and a discriminator. All of the network architectures are built with plain convolution, batch normalization, and ReLU activation layers only. The architecture details of the encoder are shown in Table 3. For an input radiography image (sized of 128×128), we first divide it into 2×2 non-overlapping patches (sized of 64×64). The encoder then extracts the patch features.

As mentioned in §3.1, the student and teacher generators were constructed identically. The only difference is that additional Memory Matrices are placed in the student generator. The architecture details of the student generator are shown in Table 4. Skip connections from the encoder are only enabled at such levels that Memory Matrices are used. After the last Memory Matrix, the non-overlapping patches are put back as a whole for further reconstruction.

As shown in Table 5, the discriminator was constructed in a more lightweight style. Note that the images are discriminated at their full resolution (*i.e.* 128×128) rather than in patches.

Table 3. Encoder structure in SQUID.

Level	#Channels	Resolution
Input	1	$(2 \times 2) \times (64 \times 64)$
1	32	$(2 \times 2) \times (32 \times 32)$
2	64	$(2 \times 2) \times (16 \times 16)$
3	128	$(2 \times 2) \times (8 \times 8)$
4	256	$(2 \times 2) \times (4 \times 4)$

Table 4. Student and teacher generator structures in SQUID. S&M denotes the usage of skip connections and Memory Matrix. Note that there is no Memory Matrix placed in the teacher generator.

Level	#Channels	*w/* S&M	Resolution
4	256	✓	$(2 \times 2) \times (4 \times 4)$
3	128	✓	$(2 \times 2) \times (8 \times 8)$
2	64		32×32
1	32		64×64
Output	1		128×128

Table 5. Discriminator structure in SQUID.

Level	#Channels	Resolution
Input	1	128×128
1	16	64×64
2	32	32×32
3	64	16×16
4	128	8×8
5	128	4×4
Output	1	1×1

B. Additional Results

B.1. Extensive Ablation Studies

In this section, we ablate three components in SQUID to fully validate their necessity and effectiveness.

Table 6. The extensive results indicate that all proposed techniques in SQUID are essential for a high overall performance.

Method	AUC (%)	Acc (%)	F1 (%)
Convolution Layers	76.9±3.3	74.2±3.3	80.7±2.7
Transformer Layers (Δ)	↑10.7	↑6.1	↑4.0
Soft Masked Shortcut	79.7±3.4	76.1±2.7	80.7±2.3
Hard Masked Shortcut (Δ)	↑7.9	↑4.2	↑4.0
Pixel-level In-painting	79.1±0.4	74.4±1.6	81.3±0.9
Feature-level In-painting (Δ)	↑8.5	↑5.9	↑3.4
Full SQUID	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**

(1) Convolutional vs. Transformer Layers: In our proposed in-painting block, a transformer layer is used to aggregate the encoder extracted patch features, and the Memory Queue 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 convolutional layer while preserving other structures.

(2) Soft vs. Hard Masked Shortcuts: In our proposed masked shortcut, skipped and in-painted features are aggregated using a binary gating mask. The intuitive question is whether such "hard" gating is necessary and a weighted "soft" addition can also achieve comparable results. To this end, instead of following Eq. 2, we conducted experiments by aggregating the patch features $\mathcal{F}$ through:

$$\mathcal{F}' = (1 - \rho) \cdot \mathcal{F} + \rho \cdot \texttt{inpaint}(\mathcal{F}), \tag{3}$$

where ρ was set to 95%, same as the best setting in SQUID.

(3) Pixel-level vs. Feature-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 image level [41, 54, 87]. 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 SQUID in-paint the distorted input images. All other settings and objective functions remain unchanged.

Summary: The results of the above three additional ablative experiments are presented in Table 6. Without using the transformer layer, masked shortcut, and feature-level in-painting as proposed, the AUC, Acc, and F1 scores decreased by at least 8%, 4%, and 3%, respectively, compared with the full SQUID setting.

B.2. Patch-MemAE

MemAE [17] with Memory Matrix is the primary baseline that we considered in this work. To further verify the effectiveness of our proposed space-aware setting, we trained additional MemAE models on patches segmented from different spatial location of input images. These multiple space-specific models were trained separately with

Table 7. We apply space-specific strategy to one of the strongest counterparts (MemAE [17]). In addition, the ensemble of spatial-aware models demands a *higher* degree of computational costs (4× more than ours), while our work proposed to encode this spatial information into the feature dictionary, ultimately requiring only one model—its efficiency is pronounced.

Method	AUC (%)	Acc (%)	F1 (%)
MemAE [17]	77.8±1.4	56.5±1.1	82.6±0.9
Patch-MemAE (Δ)	↑0.5	↑18.5	↓1.3
Full SQUID	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**

their unique space-specific patches and were then evaluated through an ensemble style to compare with our SQUID. The results are reported in Table 7.

The results of the this experiment indicate that although improvements can be observed on AUC and Acc, such space-specific ensemble upgrade still performs inferior than SQUID. Moreover, we found such ensemble of models demands a much *higher* degree of computational costs (*4*× more than ours), while in our work, we proposed to encode this spatial information into the feature dictionary, ultimately requiring only one model. Both effectiveness and efficiency are pronounced.

C. Creating DigitAnatomy

The pseudocode of creating our new benchmark dataset (DigitAnatomy in §4.1) is provided in Algorithm 1. In practice, we have implemented the algorithm into an off-the-shelf data loader that can be amended to many other different datasets (*e.g.* SVHN, CIFAR, ImageNet).

Algorithm 1 Creating DigitAnatomy

```
# a function to pick random digit instances
def pick_random(class, single_digits):
    # random pick an image with size: [28, 28]
    pick_digit = random.choice(single_digits[class])
    return pick_digit


# load MNIST digits with shape: [10, 1000, 28, 28]
single_digits = load_MNIST()

# all possible conditions
conditions = ['normal', 'missing',\
              'misorder', 'flipped', 'novel']

output = torch.zeros(3, 28, 3, 28)

# loop over digit 1-9 in order
for idx in range(1,10):

  # randomly pick a condition
  condition = random.choice(conditions)

  if condition == 'normal':
    digit = pick_random(idx, single_digits)
  # anatomy of missing digit
  elif condition == 'missing':
    digit = torch.zeros(28,28)
  # anatomy of disorder digit
  elif condition == 'misorder':
    ridx = randrom.randint(1,10)
    digit = pick_random(ridx, single_digits)
  # anatomy of flipped digit
  elif condition == 'flipped':
    digit = pick_random(idx, single_digits)
    digit = digit[::-1,::-1]
  # anatomy of novel digit
  elif condition == 'novel':
    digit = pick_random(0, single_digits)

  output[idx // 3, :, idx % 3, :] = digit

# combine all patches together
output = output.view(28 * 3, 28 * 3)
```

D. Visualization Results

D.1. Visualizations on DigitAnatomy

More reconstruction results of SQUID and the compared methods [1, 17, 61] are shown in Figure 11. Our observations from these additional results are aligned with the ones discussed in §5.1. SQUID can capture *every* appearing anomaly (highlighted in light blue) in the images and augment them back to the normal closest forms. On the contrary, although MemAE restores the normal digits the best, it is limited in detecting a few anomaly types (*e.g.* misordered and missing digits). Ganomaly is not able to perfectly recover the normal digits and also cannot generate meaningful reconstructions on the abnormal ones. f-AnoGAN, on the other hand, memorizes and generates an exemplary normal pattern that fails to respond to different inputs.

D.2. Visualizations on Chest Radiography

Figure 12 and Figure 13 show more reconstruction results of our SQUID on the ZhangLab Chest X-ray and Stanford CheXpert datasets. We observed that our method is capable of translating the input image to its "normal" counterpart and assigning larger anomaly scores to abnormal cases.

When inputting normal images, SQUID will try to reconstruct the inputs as well as possible. Due to the usage of memory modules, our framework could hardly degenerate to function as an identity mapping from inputs to outputs. Therefore, the reconstruction of normal inputs cannot perfectly recover every single detail.

When inputting abnormal images, SQUID will make larger impacts by combining previously seen normal features together into such abnormal ones. Since the generator is not trained on such hybrid features, the reconstruction results could demonstrate more obvious artifacts and blurs.

After our framework converges, the optimized discriminator can perceptually capture such inconsistencies between reconstructed normal and abnormal images and achieve anomaly detection.

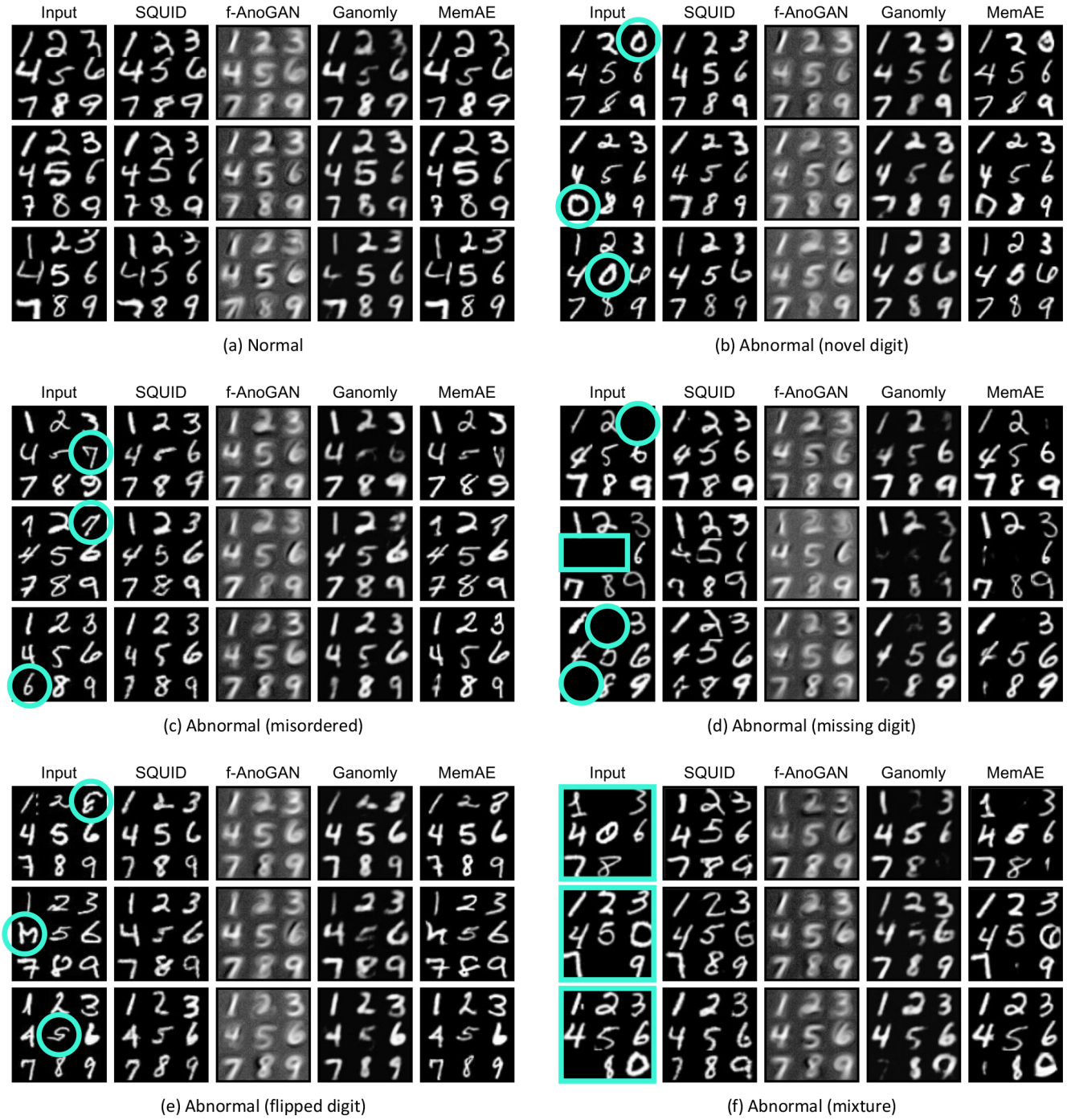


Figure 11. Comparisons of reconstruction results on DigitAnatomy of our SQUID, f-AnoGAN [61], Ganomaly [1], and MemAE [17]. Anomalies are highlighted in light blue.

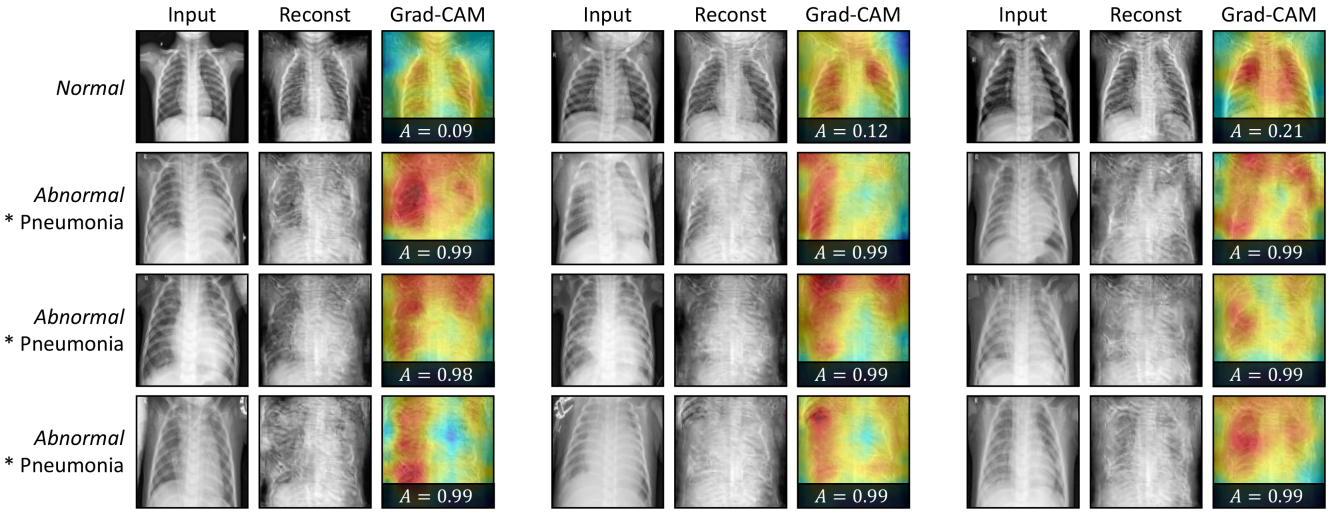


Figure 12. Reconstruction results of SQUID on the ZhangLab Chest X-ray dataset. The corresponding Grad-CAM heatmaps along with anomaly scores are shown as well.

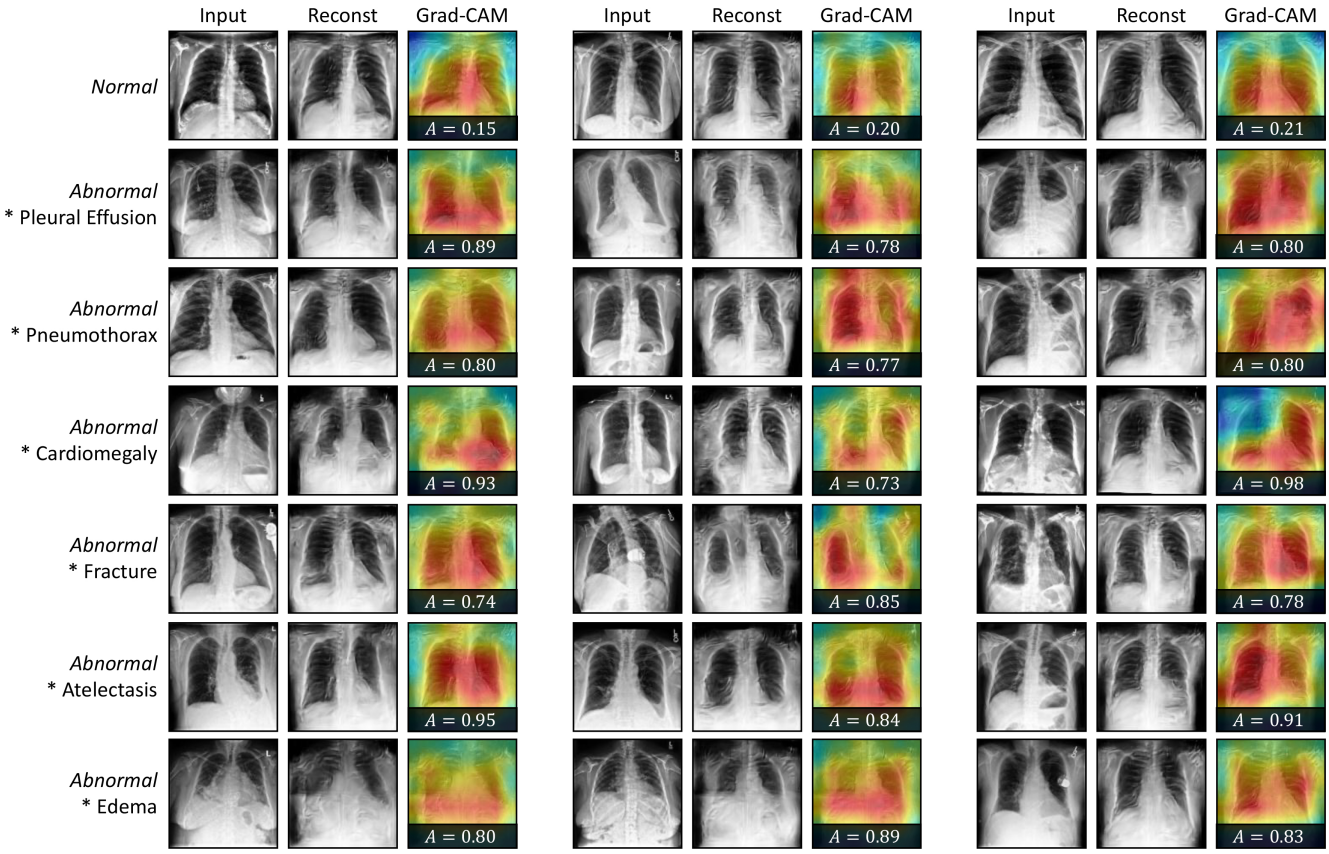


Figure 13. Reconstruction results of SQUID on the Stanford CheXpert dataset. Different disease types are separated into different rows. The corresponding Grad-CAM heatmaps along with anomaly scores are shown as well.