Deep Feature In-painting for Unsupervised Anomaly Detection in Radiography Images

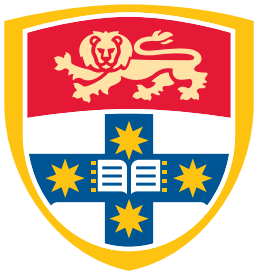
TIANGE XIANG
SID: 470082274

Supervisor: A/Prof. Weidong Cai

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Computer Science and Technology (Advanced) (Honours)

School of Computer Science
The University of Sydney
Australia

29 May 2022



THE UNIVERSITY OF
SYDNEY

Student Plagiarism: Compliance Statement

I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Tiange Xiang

Signature:  **Date**: **2022/05/29**

Abstract

Abnormal event detection in visual contents has broad application prompts in human society. Without particular human intervention, building an automatic detector that is able to spot such abnormal cases in videos or images is of great interests in the artificial intelligence and computer vision communities. Given the unknown number of possible unseen events that may exist in the dataset, the mainstream supervised learning approaches become unpractical and deep learning algorithms are desired to optimize themselves in an unsupervised manner for the anomaly detection task.

As one of the most commonly used data type in medical images, radiography images such as chest X-rays, contain detailed inner structures of ones body, which provide visual indications for anomalies. There were many research works done recently that targeted at capturing such anomalies automatically through novel unsupervised deep learning algorithms.

In this work, we achieve unsupervised anomaly detection for radiography images through: *(i)* A task reformulation to feature-level in-painting. A novel in-painting block is introduced that aggregates inter-patch contextual information; *(ii)* An upgrade of the existing Memory Matrix approach that caters to the unique characteristics of radiography images. Our design copies patch-wise features directly into the memory module for faster and better feature representations; *(iii)* The designs of the Gumbel shrinkage, masked shortcuts, and anomaly discrimination, which have never been explored in the UAD domain. We integrate all of these innovations together into a hybrid framework; *(iv)* The creation of a new dataset: DigitAnatomy, which resembles the unique properties of radiography images with easy-to-interpret handwritten digits. This dataset can help with the development, evaluation, and interpretation of anomaly detection algorithms especially for radiography imaging.

In empirical experiments, SQUID was examined against 13 state-of-the-art counterparts in unsupervised anomaly detection by a considerable margin with over 5%/8%/3% AUC improvements on three publicly released chest X-ray datasets. Our source code is available at: https://anonymous.4open.science/r/SQUID-public-3361/

Acknowledgements

First, I would like to thank my supervisor **A/Prof. Weidong Cai** for offering me the chance to start my research journey. Thanks for his kind mentoring, meticulous supervision, warm encouragement, and tireless support throughout my entire undergraduate study. I would also like to thank **Chaoyi Zhang** for his mentoring throughout my early research career and this honors thesis study, providing me with detailed guidance and tutoring.

Then, I would like to thank other research staff, research students, and collaborators in A/Prof. Weidong Cai's research group for their encouraging discussion and help. I would also like to thank **Dr. Zongwei Zhou** and **Prof. Alan Yuille** at the Computational Cognition, Vision, and Learning (CCVL) research group, Johns Hopkins University for their assistance in this joint research work that partially make up this thesis.

Lastly, I would like to particularly thank my parents **Ping Xiang** and **Zaojun Qin** for their continuous supports even in my hardest time. Thank **Liao Chen** for her trust and company.

List of Publications

Related Publications: This thesis is in part based on the following submitted research works:

- **Tiange Xiang***, Yixiao Zhang*, Yongyi Lu, Alan Yuille, Chaoyi Zhang, Weidong Cai, Zongwei Zhou, “Feature-level In-painting for Unsupervised Anomaly Detection in Radiography Images", Submitted to ***Medical Image Analysis***, 2022. (Under Review)

- **Tiange Xiang**, Yixiao Zhang, Yongyi Lu, Alan Yuille, Chaoyi Zhang, Weidong Cai, Zongwei Zhou, “In-painting Radiography Images for Unsupervised Anomaly Detection", Submitted to *The 17th European Conference on Computer Vision (**ECCV 2022**)*, 2022. (Under Review)

Other Publications:

- **Tiange Xiang**, Chaoyi Zhang, Xinyi Wang, Yang Song, Dongnan Liu, Huang Huang, Weidong Cai, “Towards bi-directional skip connections in encoder-decoder architectures and beyond", ***Medical Image Analysis***, Vol.78, p102420, 2022.

- **Tiange Xiang**, Yang Song, Chaoyi Zhang, Dongnan Liu, Mei Chen, Fan Zhang, Heng Huang, Lauren O’Donnell, Weidong Cai, “DSNet: A Weakly-Supervised Dual-Stream Framework for Effective Gigapixel Pathology Image Analysis", ***IEEE Transactions on Medical Imaging***, 2022.

- **Tiange Xiang**, Chaoyi Zhang, Yang Song, Jianhui Yu, Weidong Cai, “Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis", *2021 IEEE/CVF International Conference on Computer Vision (**ICCV 2021**)*, pp915-924, 2021.

- Xinyi Wang*, **Tiange Xiang***, Chaoyi Zhang, Yang Song, Dongnan Liu, Heng Huang, Weidong Cai, “BiX-NAS: Searching Efficient Bi-directional Architectures for Medical Image Segmentation", *The 24th Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention (**MICCAI 2021**)*,

Lecture Notes in Computer Science, LNCS Vol.12901, pp229-238, 2021.

- **Tiange Xiang**, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, Weidong Cai, “BiO-Net: Learning Recurrent Bidirectional Connections for Encoder-Decoder Architecture", *The 23th Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention (**MICCAI 2020**)*, Lecture Notes in Computer Science, LNCS Vol.12261, pp74-78, 2020.

* indicates equal first-author contributions.

A full list of publications can be found at: https://scholar.google.com/citations?user=sskixKkAAAAJ&hl

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

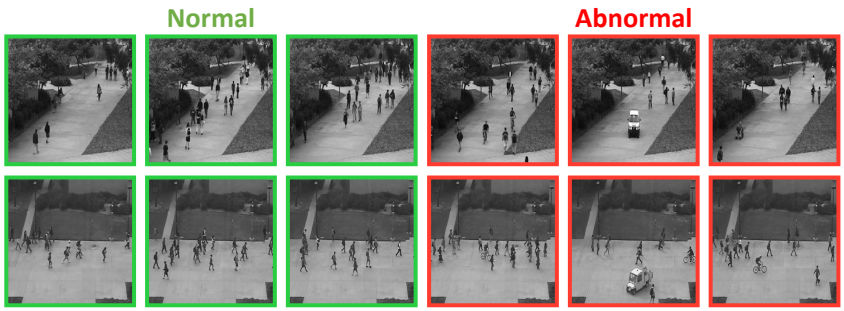
1.1 Background and Problem Definition

Artificial intelligence and deep learning algorithms have brought significant impacts to everyone's daily lives. Such algorithms have assisted people in a broad range of activities covering from general data processing all the way to life saving. As a recent trend in applied deep learning research, more and more algorithms were developed toward the medical and pathological domain. However, vision tasks in natural images and medical images are different. As an instance, when identifying objects in natural contents, their appearances locating in the images are generally not important — a cat is a cat no matter which position it appears in the image.

Relative location and orientation of body structures shown in the images are crucial factors for radiography imaging analysis. Such characteristics assist verification of normal anatomy and pathological conditions [19]. Unlike natural images that are usually captured through hand-held mobile phones with uncertain physical poses, radiography images are acquisited through strict imaging equipments and protocols, which makes the generated images in a fairly consistent viewing angle (see examples in Figure 1.1 (d)). The consistent anatomy pattern facilitates the analysis of radiography images and should be considered as an unique and significant advantage over natural images.

1.2 Current Progress and Challenges

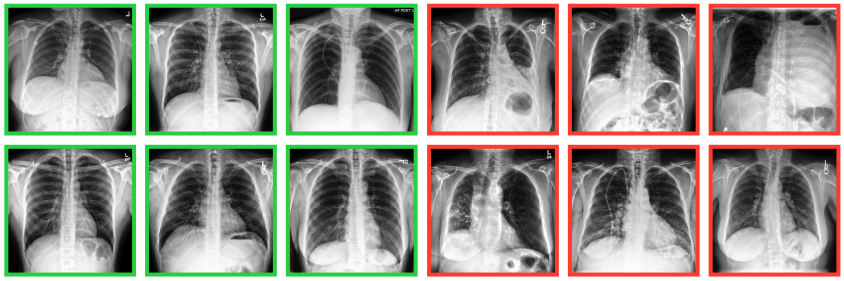
In the recent years, there are multiple works that studied the benefits of this kind of prior knowledge in enhancing the performance of deep learning models by incorporating positional information, manipulating training objective functions, and constraining landmark coordinates [3, 50, 51, 74, 91]. Based on the existing methods, in this work, we seek to answer this critical question of Unsupervised Anomaly Detection (UAD) in radiography images: *Can we exploit consistent anatomical patterns and their*



(a) Anomaly Detection in Crowded Scenes

(b) Anomaly Detection in Textures and Objects

(c) Anomaly Detection in Digit Anatomy



(d) Anomaly Detection in Chest Anatomy

FIGURE 1.1. The form of anomaly is different in photographic images and radiography images, which makes anomaly detection in radiography images a new challenge. Compared to natural images, chest X-rays are more spatially structured because of consistent imaging protocols. However, anomalies in radiography images are subtle and harder to be detected.

spatial information to strengthen deep learning models in detecting anomalies from radiography images without relying on ground truth annotations?

We approach this challenge by formulating anomaly detection as an in-painting task to adhere the consistency in objects' semantics and layouts. To be more specific, we propose **S**pace-aware memory **QU**eues for **I**n-painting and **D**etecting anomalies in radiography images (abbreviated as SQUID). When training on normal-only images, SQUID maintains a *dictionary* that is automatically refreshed with extracted training features in space-aware anatomical patterns. Attribute to the anatomy consistency, the same scanning of body part usually reveals very close visual patterns, which constrains the total number of unique regions. A similar feature pattern is then fetched from the dictionary to be decoded back into the pixel space. The whole framework is self-supervised following the reconstruction loss.

Since anomaly patterns have not been seen by the dictionary, reconstruction of abnormal images is expected to be unrealistic. As a result, SQUID can identify the anomaly by discriminating the quality of the in-painting task. The success of SQUID relies on two assumptions [94]: *(i)* Anomalies only occur rarely in training data; *(ii)* Anomalies differ from normal patterns significantly. Consequently, the dictionary will reflect a general distribution of anatomical patterns in the normal human anatomy.

We conducted experiments on three large-scale, publicly available radiography imaging datasets to verify the effectiveness of our method: the ZhangLab dataset [34], the CheXpert dataset [30], and the COVIDx dataset [82]. SQUID achieved the best results compared to state of the art on all of the datasets. The qualitative visualizations clearly demonstrate SQUID's superiority over the existing methods. In addition, we built a novel dataset (DigitAnatomy) to better demonstrate the attributes of spatial associations and shape consistencies in radiography images with easy-to-interpret hand-written digits only. Our goal of DigitAnatomy is to make the development, evaluation, and interpretation of anomaly detection algorithms for chest X-rays as simple as possible.

1.3 Main Contributions

In summary, our major contributions in this work are six-fold:

(1) We reformulate the unsupervised anomaly detection for radiography images as a feature-level in-painting task. A novel in-painting block is introduced to achieve inter-patch in-painting by using a Transformer layer [80].

(2) We propose the Space-aware Memory Queue that upgrades the existing Memory Matrix approach to cater to the unique characteristics of radiography images. Our upgrade copies patch-wise features directly into the memory module for faster and better feature representations.

(3) We design the functional modules: Gumbel shrinkage, masked shortcuts, anomaly discrimination that have never been explored in the UAD domain. We integrate all of these innovations together into a hybrid framework—SQUID.

(4) We create a new dataset, namely DigitAnatomy that uses handwritten digits to resemble the unique properties of radiography images. We hope this dataset can help the development, evaluation, and interpretation of anomaly detection algorithms in the relevant communities.

(5) We examine SQUID on 3 challenging public benchmarks against 13 existing unsupervised anomaly detection methods. Impacts of the proposed modules along with the choices of hyper-parameters are exhaustively studied as well.

(6) We, for the first time, investigate the robustness of UAD methods to the normal/abnormal ratio in the training set. This study relaxes the rigid disease-free training protocol and verifies UAD methods in a strict unsupervised setup.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter 2, we review classic and recent advances on the topics that are mostly close to this work: general deep learning networks, general anomaly detection, unsupervised anomaly detection for medical images, and memory networks. In Chapter 3, we present our main algorithm: SQUID for unsupervised anomaly detection in radiography images. For each of the newly introduced components, we articulate their motivations and design details. In Chapter 4, we state our experimental designs: evaluation benchmarks; comparing methods; metrics; protocols; implementation details. We also demonstrate major results obtained during the experiments and present extensive studies to fully validate our method. In Chapter 5, we discuss current limitations of SQUID and potential future research directions regarding of the limitations. Finally, in Chapter 6, we conclude this thesis with summaries and reflections.

CHAPTER 2

Literature Review

2.1 Deep Learning Networks and Frameworks

Deep learning algorithms are powerful approximators for complex non-linear equations. Many real-life applications: image recognition [24], action detection [81], autonomous driving [29], content generation [33] benefit from deep learning and self-optimized neural networks. Among the large coverage of deep learning based research works, in this section we mainly focus on reviewing the ones that are important and contain necessary prior knowledge to comprehend the proposed method in this work.

ResNet [24] is one of the most influential works in the machine learning community. It starts a new era for deep learning by introducing the very first deep architectures with over 1000 layers. Before ResNet, deep networks are believed to be problematic due to unstable gradients and performance degradation. More layers lead to larger variance in the intermediate feature distributions. When calculating layer-wise gradients, the partial derivatives could easily converge to zero or escalate to very large numbers. Therefore, most of prior networks were built with very few layers [37, 72]. In order to optimize deep neural networks, He *etal.* proposed to link a stack of convolutional layers with feature shortcuts (through element-wise summation) that enables the network to learn feature residuals. There are two major benefits of such a simple design: *(i)* Intermediate feature distribution gap can be bridged by merging features at different layers; *(ii)* The gradients calculated beforehand can be identically preserved through the shortcuts for better gradient stability. To further accelerate deep network optimization and reduce the overall memory usage, the authors proposed to use point-wise convolutions with 1×1 kernels to create channel-wise bottlenecks to compress intermediate feature maps. This design inspires various follow-up works including SQUID: we also adopt a similar bottleneck design as part of our in-painting block.

ResNet was originally proposed as a backbone network for image recognition. Concurrently, another pioneering work: U-Net [59] was proposed to advance image segmentation. Early neural network based

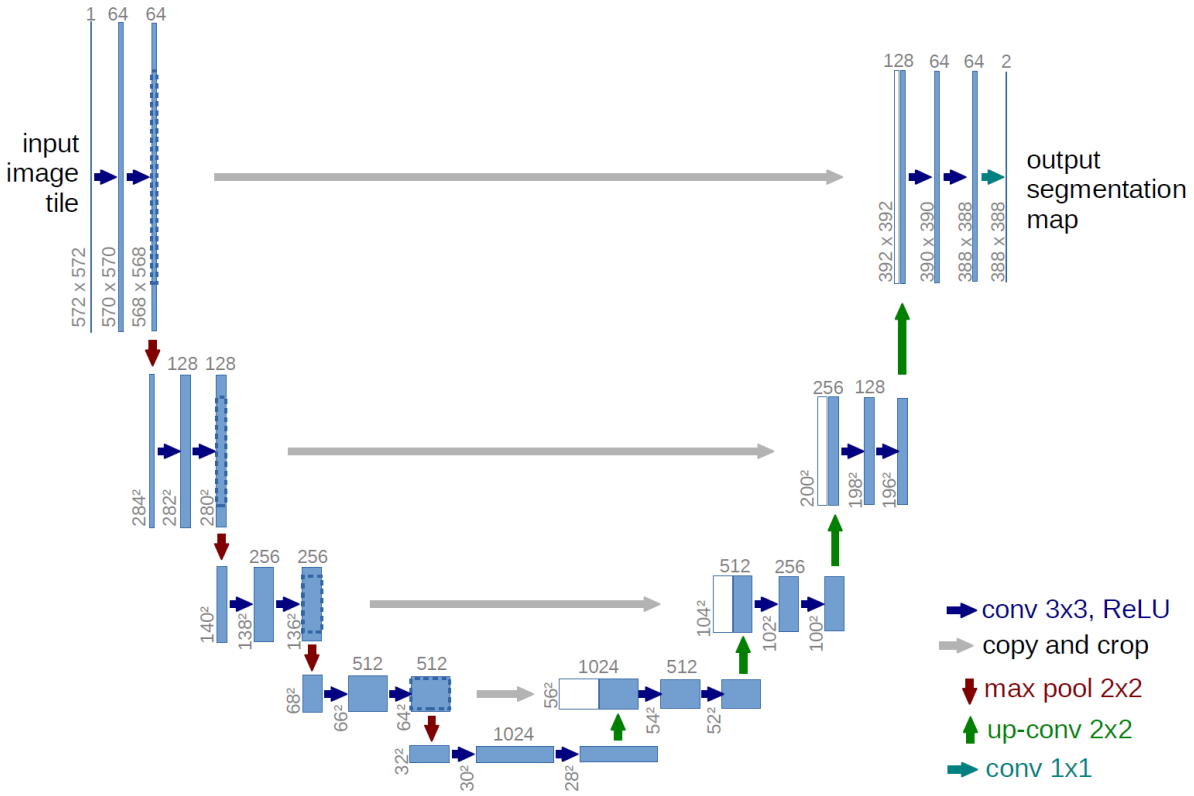


FIGURE 2.1. U-Net network structure [59].

image segmentation algorithms followed the design of Auto-Encoder with paired encoder and decoder network. U-Net improves the classic structure with skip connections that ship encoded features to the decoder at each level. An overview of the proposed structure is outlined in Figure 2.1. With the same philosophy as ResNet, the skip connections supplement the decoder with semantics obtained at the encoder for better feature aggregations. Moreover, gradients passed from the decoder can be well preserved to the encoder for more stable back propagation. This simple network was proved to be effective on a broad range of tasks [4, 84]. In SQUID, we utilized this simple architecture as our backbone. We believe our method, with only the most basic networks, is already capable of achieving superior anomaly detection performances.

Another important work that is closely associated with our method is the Generative Adversarial Nets (GANs) [17]. Unlike the traditional gradient descent based optimization strategy, GANs utilize adversarial learning to optimize a joint of two networks that are concurrently trained to compete against each other. The two networks are termed as the generator and discriminator. The generator is given with

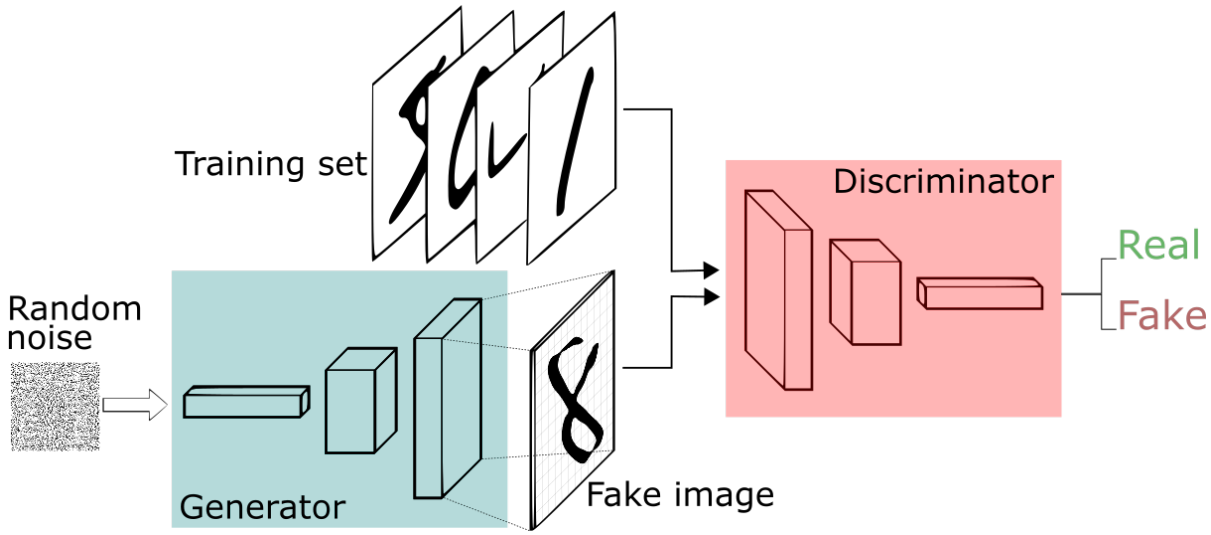


FIGURE 2.2. Framework of Generative Adversarial Nets [17].

vectors filled with random values and is expected to generate as realistic images as possible. The discriminator, on the hand, criticizes on the generated images and tries to discriminate whether an image is real or fake. In an ideal scenario, the two-network system is optimized adversarially to reach the Nash equilibrium. At the end, the optimized generator can be used alone to generate fake data in high qualities by giving any random vectors. An overview of the GAN framework is presented in Figure 2.2. This generation framework is especially helpful to forge data that fit a particular distribution. In unsupervised anomaly detection, network is expected to generate fake normal images of any given inputs, so that the disparity between the generation results and the inputs can be assessed and captured. We therefore adopt the GAN framework as the primary protocol in SQUID.

2.2 Anomaly Detection in Natural Imaging

The objective of anomaly detection is to find uncommon events that are away from the regular data distribution [54]. The history of machine learning based anomaly detection can be traced back to the last century when learning algorithms were just born. One of the most early attempts was to train a Support Vector Machine (SVM) [67] with only one-class targets. The simple classifier will output uncommon logits whenever an anomaly data point is showing. Another milestone approach was based on dictionary learning, where the authors proposed the Sparse Reconstruction Cost (SRC) [9] over the normal dictionary to measure how normal the testing data is. The algorithm was further enhanced by a selection criteria to meet the sparsity consistency constraint.

With the thriving of deep learning, most later works approach anomaly detection as an unsupervised learning problem due to insufficient representations of anomalies [11, 25, 26, 40, 41, 46, 60, 71, 95]. Generally, deep learning based methods can be classified into two major categories: reconstruction-based and density-based. Reconstruction-based methods build Auto-Encoder like neural networks to learn image self-reconstructions without using explicit supervisions [7, 70, 76, 92, 93]. When learning on normal data only, the network decoder could easily capture the consistency across normal patterns. In this way, potential anomalies can be found by assessing reconstruction errors between original inputs and their reconstructed results. Density-based approaches, on the other hand, follow the variational inference framework to build Variational Auto-Encoders (VAEs) [2, 36, 64]. VAEs detect anomalies by estimating a statistical model representing the normal data. All of the existing methods, however, cannot explain the possible irregularities when using the learned distribution for normal images. Considering this, we address these problems by following the dictionary learning approach that maintains an explicit dictionary collecting visual patterns of homogeneous data samples in the semantic feature space.

In a recent work, CutPaste [43] was proposed as an augmentation technique in the input distribution to forge 'abnormal images' and transform the unsupervised learning problem into a self-supervised learning task. Specifically, normal images are binary masked by random patches or 'scars' that are copied from another area in the same image. With such manually simulated anomaly, neural networks can be trained to make binary classification based on the augmentation. Figure 2.3 shows the training pipeline of CutPaste. Although been verified through a wide range of empirical experiments, it is clear that the CutPaste augmentation pattern can not generalize well to all possible anomaly types and may fail on certain cases apart from structural issues.

Instead of simulating anomalies, is it possible to utilize networks that are pre-trained on large-scale datasets of more anomaly occurrences? PANDA [58] studied such a feature adaptation schema. The authors experimented on utilizing pre-trained features from other tasks (e.g. semantic segmentation) for anomaly detection and found such a simple adaption already beat state-of-the-art methods that were trained from scratch. Based on this finding, the authors designed multiple feature adaption techniques on the one-class classification setting. However, pre-trained models are hard to obtain for uncommon data domains, which makes the proposed method impractical for medical images.

Without accessible pre-trained models, feature distillation can be used as a surrogate solution. M-KD [62], on the other hand, adopted the distillation paradigm to address two common issues in anomaly detection: *(i)* Limited data samples are insufficient to learn generalizable models on their own; *(ii)*

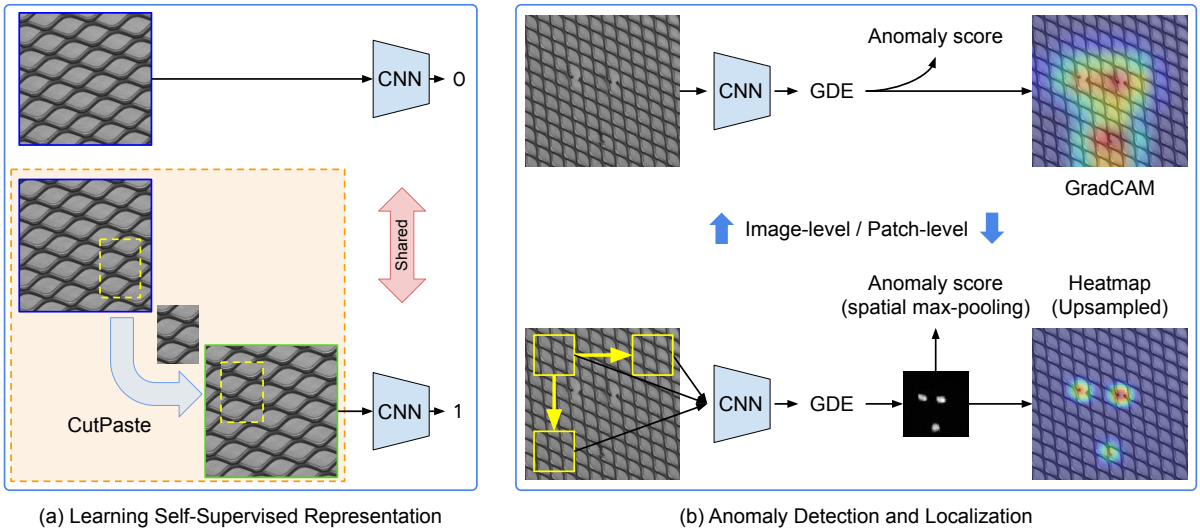


FIGURE 2.3. CutPaste training pipeline [43].

Without pre-assumptions on the data distribution, mixed training of both normal and abnormal data can confuse model learning. Inspired by them, our SQUID also adopts a distillation strategy. However, we utilize such a structure to distillate input-aware features only and we completely disable the teacher network during inference.

Noteworthy, there are also several other works investigated how image in-painting may assist anomaly detection. In a typical paradigm, random pixels of the input image are masked out and the model is trained to recover the missing parts in a self-supervised way [22, 53, 88]. Unfortunately, visual distortions exist in the raw pixels can disrupt network optimization and undermine overall performances.

2.3 Anomaly Detection in Medical Imaging

Comparing to the efforts made on natural images, explorations on medical domain are relatively insufficient [66]. In a general paradigm, deep learning models are trained with normal images only trying to overfit the network. After training, such a network will perform very well on normal image reconstruction but fails to recover anomaly regions that have not seen during training.

One of the classic algorithms followed this paradigm is AnoGAN [65]. This method was built based on a GAN with two separate networks: a generator network and a discriminator network. During training, the generator is given random vectors as inputs and is optimized to generate high-resolution images that 'look like' the training (normal) images perceptually. Without any pixel-to-pixel supervisions, the

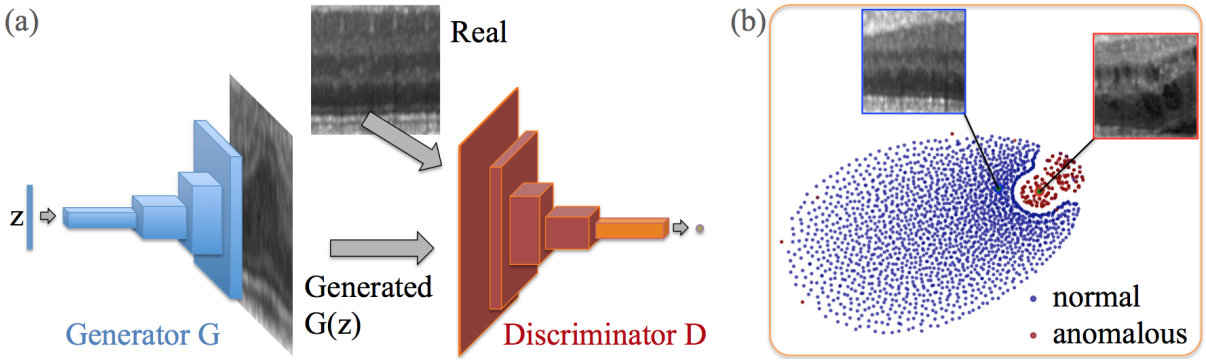


FIGURE 2.4. The framework of AnoGAN [65].

generator is updated in accordance to the discriminator in an adversarial learning style: the gradients calculated for the discriminator are also flowed to the generator for network optimization. When the training converges, the discriminator is then able to distinguish between the appearance of normal images and unseen abnormal ones. The AnoGAN framework is briefly outlined in Figure 2.4.

As a direct upgrade of AnoGAN, f-AnoGAN [64] was proposed to speed up network training. In AnoGAN, vectors of random values are used as inputs for the generator. It is practically hard for the network to fit on random patterns to generate high-quality images. f-AnoGAN improves this by using an extra encoder network to generate the input vectors. The authors investigated two different training protocols and both of them yield better training/inference efficiency.

In another work, Marimont *et al*. [52] designed an Auto-Encoder based neural network to learn the distribution of normal images. Unlike the generation method as in AnoGAN and f-AnoGAN, the proposed method forces the network to learn a mapping between spatial coordinates and probabilities of tissue type classification. During inference, the learned mapping is used to restore a normal image that looks most similar to the input image.

Most recently, combined with self-supervision techniques, SALAD [90] was designed as a hybrid framework to learn uncommon patterns explicitly through manual speculated anomalies similar to CutPaste. Specifically, during training, normal images are injected with artificial anomalies via pixel corruption and pixel shuffling. The forged abnormal images are used as extra training data to supervise the training of a GAN. According to the experimental results, such a design assists learning of robust feature representations for all kinds of patterns. However, restricted by the limited number of anomaly types, SALAD must rely on strong prior knowledge and assumptions. The framework of SALAD is briefly outlined in Figure 2.5.

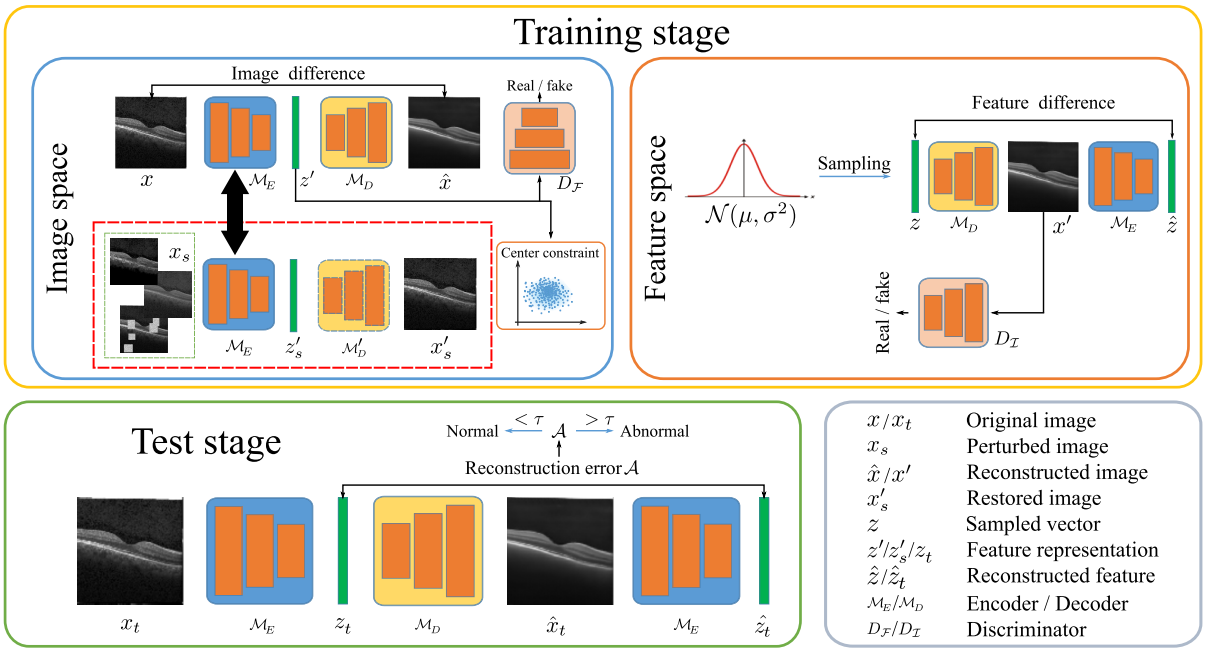


FIGURE 2.5. The framework of SALD [90].

Radiography images hold consistent anatomical patterns over photographic images, one should take this unique characteristic into method design. However, anomalies in radiography images are usually too difficult to be captured due to subtle visual clues and overlapping anatomic structures (Figure 1.1 (d)).

2.4 Memory Networks

Incorporating dictionary learning into machine learning has been explored widely during the past years [5, 14, 32, 38, 42]. Dictionaries are good at memorizing useful patterns that can be retrieved whenever needed to assist down-streaming tasks. Given the advantages of dictionary learning, researchers proposed 'memory module' for unsupervised anomaly detection.

MemAE [15] introduced the memory Auto-Encoder for anomaly detection. As shown in Figure 2.6, a learnable memory module is placed at the middle of the Auto-Encoder to fit on the encoded features. When training on normal images, the memory module will learn from normal features only for perfect normal image reconstruction. During inference, encoded features of input images are augmented by replacing them with the most similar items in the memory module. In this way, regardless of input, only learned features in the memory module are passed to the decoder for reconstruction. As a result, anomaly can be detected by assessing the reconstruction quality.

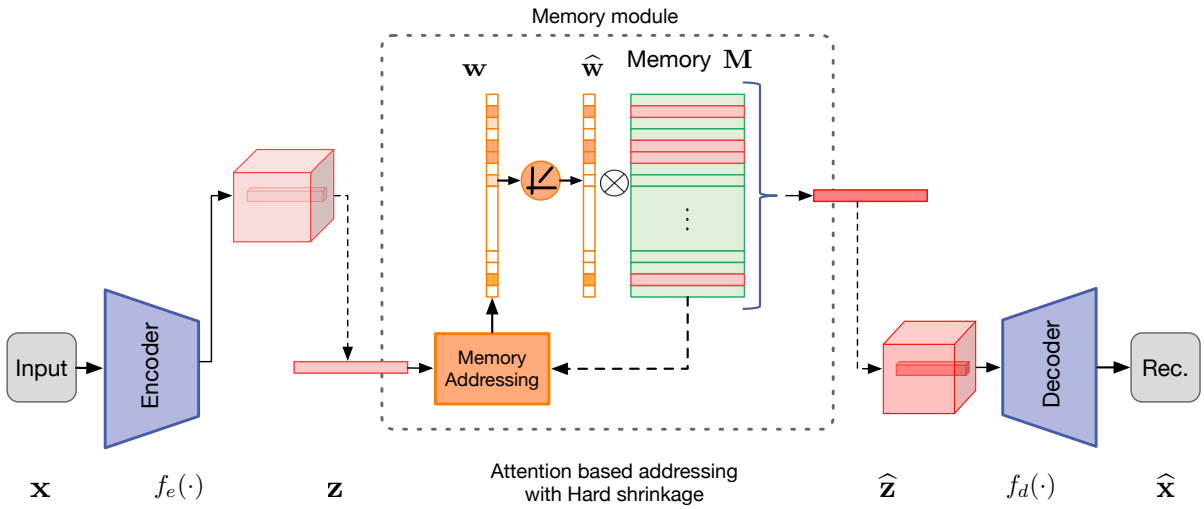


FIGURE 2.6. The network architecture of MemAE [15].

Based on this paradigm, Park *et al.* [55] proposed a more efficient memory module based method. Instead of learning the memory module along with network training, the authors proposed a set of updating rules to refresh the memory module without any gradient calculation. Two specific operations: 'read' and 'update' were designed to help the above process. There are two major advantages of using non-learnable memory module: *(i)* Training becomes more computation and memory efficient; *(ii)* Input-specific features can be better integrated into the memory.

In this paper, we advance memory module a step further by introducing the effective yet efficient **Memory Queue** for unsupervised anomaly detection particularly in radiography images.

CHAPTER 3

Methods

In this chapter, we begin with an overview of our unsupervised anomaly detection method SQUID in §3.1. We then analyze the limitations of existing Memory Matrix [15] and present our solutions by introducing the Space-aware Memory Queue in §3.2. With the proposed setup, we re-formulate anomaly detection as a task of feature-level in-painting in §3.4. §3.5 entails the details of using the discriminator for assessing the in-painting quality and alerting anomalies in the test images. Subsequently, we summarize the target functions for optimizing the entire framework in §3.6. In addition, network architecture details are presented in §3.7. Finally, in order to demonstrate the unique characteristics of chest X-rays intuitively, we design DigitAnatomy as a new benchmark in §3.8 to assist the development and interpretation of anomaly detection algorithms for radiography imaging analysis.

3.1 SQUID Overview

SQUID comprises three stages: feature extraction, image reconstruction, and anomaly discrimination.

Feature Extraction. Following a common protocol, input images are first projected into the semantic space by an encoder network. Compared to raw pixels, extracted features are in high dimensional representations with less corruptions that are easier to be processed subsequently. In a pre-processing step, we divide the input image into $N \times N$ non-overlapping patches and extract the patch features separately, which will later be used for in-painting and reconstruction. In practice, one can build the encoder network with any advanced architectures [13, 75], however to highlight on the effectiveness of the framework itself, we adopt only the most basic building layers.

Image Reconstruction. We adopt a dual-network design following the teacher-student paradigm to reconstruct the input separately. Similar to [15], a dictionary of anatomical patterns will be dynamically maintained in a queue (§3.2). To be more specific, the teacher generator aims at reconstructing the input

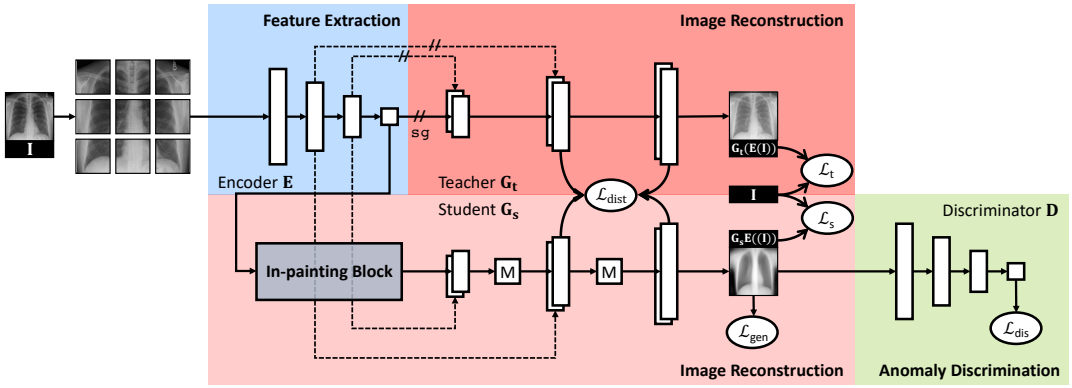


FIGURE 3.1. **SQUID framework**. There are three sequential stages: feature extraction, image reconstruction, and anomaly discrimination. **M** denotes Memory Matrix.

image with the features that are directly extracted by the encoder. The student generator, however, uses augmented features for the reconstruction—the features been processed by our in-painting block (§3.4). Our student generator will reconstruct a corresponding 'normal' image resembles the input. Only the student generator reconstructed images will be used for anomaly discrimination (§3.5). The teacher generator functions as a regularizer that prevents the student generator from mode collapse—constantly generating the same image regardless of the input. Backpropagation between the teacher generator and the encoder is disabled and we show its empirical benefit in §4.4.

Anomaly Discrimination: Inspired by adversarial learning [64, 65], we consider another discriminator network to assess the quality of reconstructions and eventually identify the generated image is real or fake. Training in the adversarial style, both the teacher and student generators will be updated by the gradients passed from the discriminator. The discriminator aims at competing the two generators and encouraging them to reconstruct realistic normal images. After fitting the whole framework on pure normal images, the generator is expected to reconstruct abnormal images in bad quality, which will then be captured by the discriminator and detected as an anomaly (§3.5).

3.2 Inventing Memory Queue as Dictionary

3.2.1 Motivation

Incorporating Memory Matrix in neural networks for anomaly detection was first proposed in MemAE [15]. Since then, it has been widely studied and extended in broader areas [16, 48, 87]. To counterfeit normal features, the Memory Matrix *augments* input features by assembling a similar pattern in the

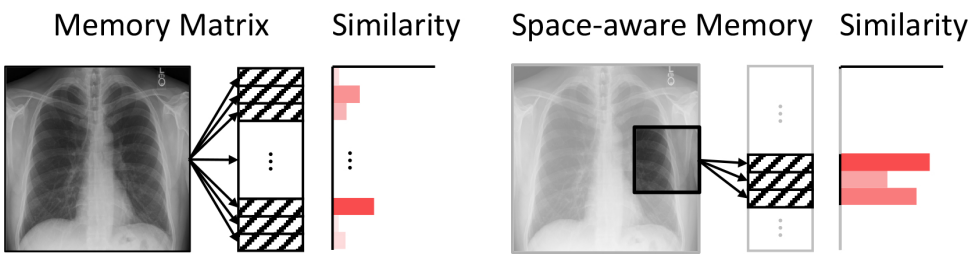


FIGURE 3.2. **Space-aware Memory.** For unique encoding of location information, we restrict each patch to be only accessible by a non-overlapping region in the memory.

dictionary. In its original design, features of the entire image are used for an image-wise feature matching. This operation, however, hinders the implicit spatial information embedded and is not effective for reconstructing MRIs. Considering this, essential upgrades are required for better space awareness.

3.2.2 Space-aware Memory

To this end, we present Space-aware Memory—a simple upgrade for the baseline Memory Matrix, which stores input features in separate patches rather than the whole image to preserve the spatial information. We seek to build unique relationship between the patch location and memory region. The memory matrix $\mathbf{M}$ is divided into blocks $\{\mathbf{M}_{i,j} \in \mathbb{R}^{N\times C}\}$, each associated with a patch at location (i, j), where N and C denote the number and the dimension of items, respectively. Let $\mathcal{F}_{i,j} \in \mathbb{R}^{C}$ denote the feature of patch (i, j), we obtain the augmented feature $\mathcal{N}$ as follows:

$$\mathcal{N}_{i,j} = \sum_{k=1}^{N} \mathcal{G}(\mathbf{w}^k)\mathbf{M}^k_{i,j}, \tag{3.1}$$

where $\mathbf{w}^k$ is the similarity score computed by dot product between $\mathcal{F}_{i,j}$ and the k-th memory item $\mathbf{M}^k_{i,j}$. $\mathcal{G}(\cdot)$ is the Gumbel-softmax operation (later in §3.3).

By doing so, we restrict the searching space of the Memory Matrix to only the specific memory region subjects to the patch location. In other words, to augment a patch at a particular location, only the corresponding region in the Space-aware Memory becomes accessible (Figure 3.2). We name this efficient upgrade as *Space-aware Memory* because it introduces space-wise priors to the Memory Matrix.

3.2.3 Memory Queue

Based on the space-aware setting, we upgrade the updating rule of the baseline Memory Matrix to represent features better. Note that, in the baseline, the matrix is optimized together along with the backbone

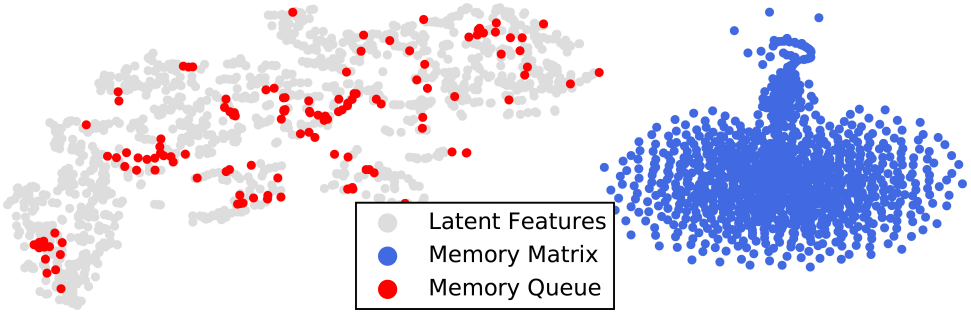


FIGURE 3.3. **Feature distribution comparisons.** t-SNE [79] visualizations of the encoded training features (gray), the learned Memory Matrix [15] (blue), and the patterns stored in our Memory Queue (red). The learned Memory Matrix deviates significantly from the distribution of encoded training features. While the features stored in our Memory Queue (as direct copies of training features) are in an identical distribution.

network during training, hence the features in the matrix are not directly generated by the encoder. Instead, the matrix itself leans to approximate feature distributions of the encoder. This training-based strategy causes distribution deviation and may lead to unstable reconstruction results.

To solve this problem, we propose to store the encoded features in a 'copy-and-paste' manner, such that the encoded features are directly copied in the matrix during training. We compare the feature distributions learned by the baseline strategy and our proposed strategy in Figure 3.3. Clearly, the learned features (blue dots) distribute dramatically different to the actual encoded training samples (gray dots). Whereas, our proposed Memory Queue shares an identical feature distribution (red dots).

However, at every training iteration, copying features into the memory module consumes unneglectable time that significantly slows down the overall training efficiency. Let's suppose there are N patterns in Memory Matrix and M training steps, the baseline storing strategy (neglecting the time cost of training) [15] demands a time complexity of $\mathcal{O}(NM)$. We designed a more efficient implementation: for each data batch, we copy the corresponding encoded features into Memory Matrix for **only once** at each training step, as outlined in Figure 3.4 (c). This simple modification yields a linear time complexity of $\mathcal{O}(M + cM)$ with the copy-and-paste operation in a constant time c. When the storage reaches its maximum, we refresh the Memory Matrix from the beginning by following the first-in-first-out (FIFO) paradigm. In this way, we call such a updated Memory Matrix as **Memory Queue**.

During training, our Memory Queue continuously copies encoded training samples into the matrix. After training, Memory Queue can be used to look up real and seen normal patterns. We query the most similar features in the Memory Queue by following Equation 3.1 as well.

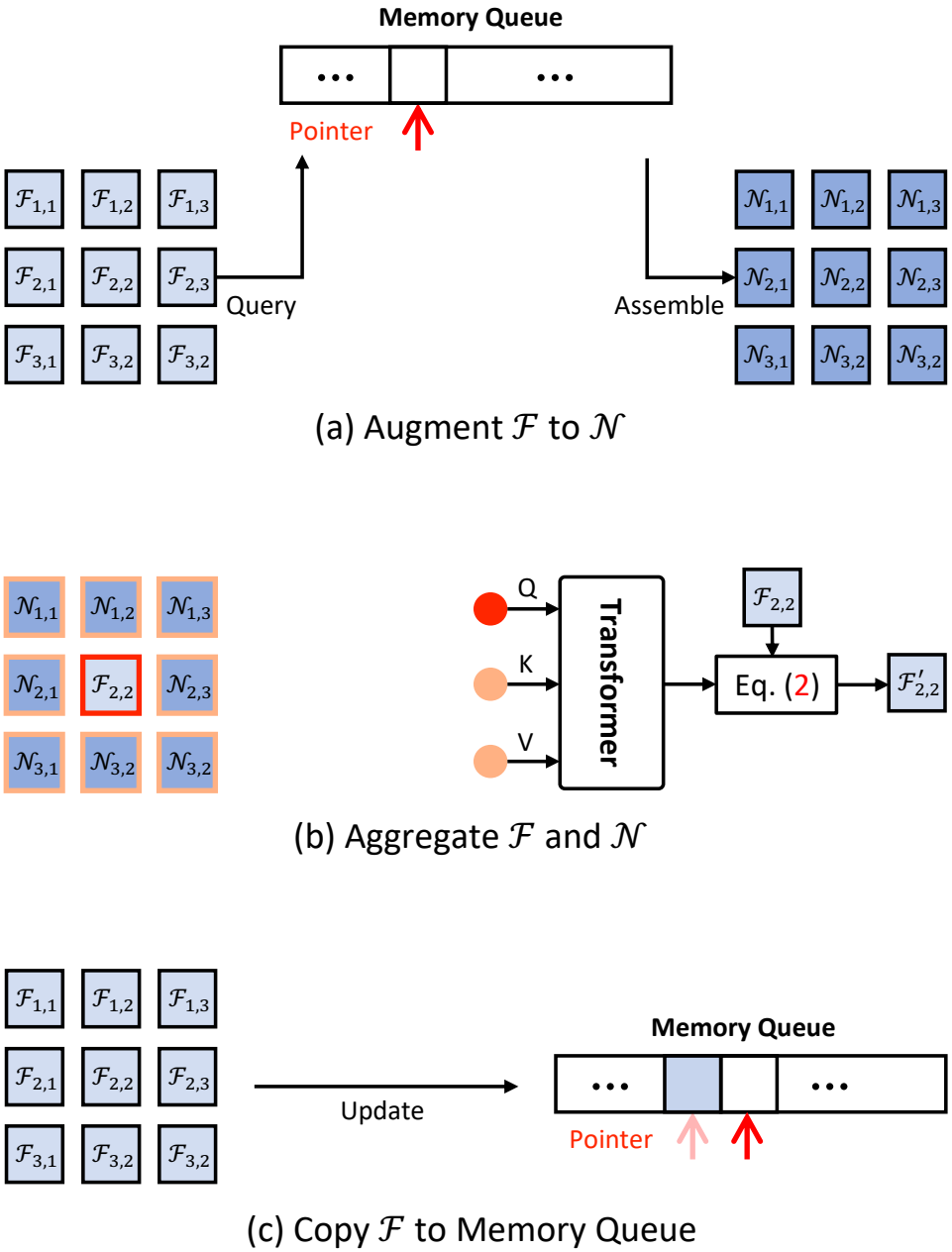


FIGURE 3.4. **In-painting block workflow.** (a) After encoding each of the non-overlapping image patches, the patch-wise features $\mathcal{F}$ are then 'augmented' by the Memory Queue: most similar items in the Memory Queue are weighted summed to assemble $\mathcal{N}$; (b) The inpainting process runs in a sliding-window manner that for each patch feature $\mathcal{F}$, its all eight neighbors $\mathcal{N}$ are used as query and key/value respectively to a Transformer layer for inpainting a 'normal' patch of features $\mathcal{F}'$; (c) In the space-aware setting, each Memory Queue region copies features $\mathcal{F}$ at corresponding spatial locations with the help of a pointer. This step is only activated during training.

3.3 Gumbel Shrinkage

3.3.1 Motivation

It has been proved that restricting memory bandwidth is beneficial for feature assembling [15, 18]. The existing methods restrict the number of memory items by clipping small similarities, which is controlled by a manually defined threshold. In other words, a similarity score $\mathbf{w}$ is set to be 0 if it falls below the threshold. However, there may exist extreme cases that the similarities of all items are significant enough exceeding the threshold or there are not similar features at all and all items are clipped to 0.

3.3.2 Combining Hard Shrinkage and Gumbel Softmax

Instead of hard thresholding, we propose to activate top-k most similar memory items at each query time. However, in the training-based Memory Matrix, only the top-k items are able to receive gradients during training and all others will not be updated as expected. Considering this, we propose the **Gumbel Shrinkage** schema inspired by [31]. During forward pass, only the top-k most similar memory items are activated to assemble 'normal' features. During backward propagation, the calculated gradients on the k items are distributed to all others in the softmax manner. Our Gumbel Shrinkage $\mathcal{G}$ is written as:

$$\mathcal{G}(\mathbf{w}) = \texttt{sg}(\texttt{hs}(\mathbf{w}, \texttt{topk}(\mathbf{w})) - \phi(\mathbf{w})) + \phi(\mathbf{w}), \tag{3.2}$$

where $\mathbf{w}$ again denotes the similarity between the input patch feature and memory items, $\texttt{sg}(\cdot)$ denotes the stop-gradient operation, $\texttt{hs}(\cdot, t)$ denotes the hard shrinkage operator with threshold t, and $\phi(\cdot)$ denotes the Softmax function. In the forward pass, Gumbel Shrinkage allows top-k most similar items in the Memory Matrix to be combined; During back propagation, Gumbel Shrinkage functions as a Softmax function. Although our Memory Queue is not learning-based, we apply Gumbel Shrinkage to both Memory Queue in the in-painting block and Memory Matrix in the student generator (§3.7).

3.4 Formulating Anomaly Detection as In-painting

3.4.1 Motivation

Originally, image in-painting [44, 56] was proposed as a concept of restoration that recovers corrupted parts in the images utilizing underlying contextual information. The in-painted pixels, on the other hand,

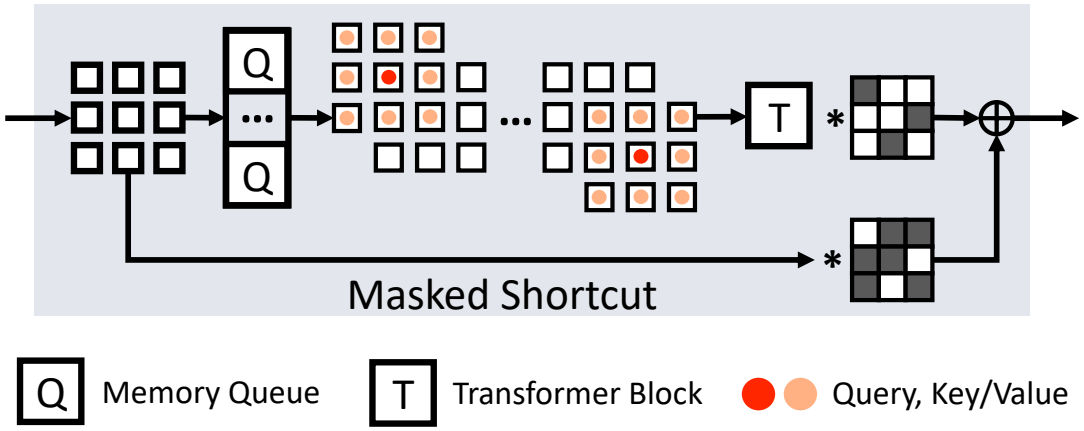


FIGURE 3.5. **In-painting block**. Patch features are augmented into their normal counterparts with the Memory Queue, a Transformer layer, and the masked shortcut.

usually include many observable defects such as boundary artifacts, distorted and blurry predictions, especially when employing neural networks [47].

When formulating anomaly detection as an image-level in-painting task, these unexpected artifacts may undermine image reconstruction quality and lower the overall detection accuracy [70, 92]. Because the discriminator could focus more on the artifacts rather than the actual abnormalities. To address this issue, we propose to perform in-painting at the feature space rather than the raw pixel space. We argue that the extracted image features are highly abstract and such latent representations are believed to be more suitable for anomaly detection since they are robust to pixel-wise noise and distortions.

3.4.2 In-painting Block

We design a novel functional block, namely *in-painting block* to perform the feature level in-painting. The block, as structured in Figure 3.5, consists of three components: our Space-aware Memory Queue, a Transformer layer [80], and a shortcut connection. Specifically, after the feature extraction stage, the obtained $w \times h$ non-overlapping patch features $\mathcal{F}_{\{(1,1),\cdots,(w,h)\}}$ are first augmented (Equation 3.1) to the most similar *normal* features $\mathcal{N}_{\{(1,1),\cdots,(w,h)\}}$ with the help of Memory Queue (Figure 3.4 (a)).

Note that the Memory Queue is filled with features of training images, hence the augmented features $\mathcal{N}$ does not attribute the current input image. To fully utilize input information, we design the in-painting process as the aggregation of both encoded features $\mathcal{F}$ and augmented features $\mathcal{N}$ using a Transformer layer [80]. We restore the central feature based on neighboring features in a sliding-window style: each patch $\mathcal{F}_{i,j}$ and its spatially adjacent eight augmented *normal* patches $\mathcal{N}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}}$ are used as

conditions to obtain the in-painted patch $\mathcal{F}'_{i,j}$ at location (i, j) (Figure 3.4 (b)). We do this by flattening $\mathcal{F}_{(i,j)} \in \mathcal{R}^{1\times *}$ to be the query token and $\mathcal{N}_{\{(i-1,j-1),\cdots,(i+1,j+1)\}} \in \mathcal{R}^{8\times *}$ to be the key/value tokens as inputs to the Transformer layer. Inspired by [24], we employ extra pair of point-wise convolutions at the start and the end of our in-painting block to create a channel-wise bottleneck for efficiency.

3.4.3 Masked Shortcut

As suggested by [24], we include a shortcut connection across the entire in-painting block to learn residual features and assist network optimization. However, since the framework target is to reconstruct the input image itself, a naive shortcut connection will significantly deteriorate the effectiveness of our in-painting block (please see empirical studies in §4.4). Borrowing the idea from [85], we use a random spatial mask to binary gate the shortcut during training. By introducing uncertainty into the shortcut, the network will be forced to learn from both in-painted and skipped features. Formally, given the encoded features $\mathcal{F}$, the output of our in-painting block is defined as:

$$\mathcal{F}' = (1 - \delta) \cdot \mathcal{F} + \delta \cdot \texttt{inpaint}(\mathcal{F}), \tag{3.3}$$

where $\texttt{inpaint}(\cdot)$ denotes our in-painting block, $\delta \sim \textit{Bernoulli}(\rho)$ denotes a Bernoulli variable with a gating probability of ρ. Note that such shortcut is only activated during training. During inference, we completely disable the shortcut for deterministic predictions:

$$\mathcal{F}' = \texttt{inpaint}(\mathcal{F}). \tag{3.4}$$

3.5 Anomaly Discrimination

3.5.1 Motivation

Recall that the in-painting block aims at augmenting patch features of any input images (either normal or abnormal) into similar *normal features* that have been seen during training. With the augmented features, our student generator reconstructs a 'normal' image back to the input pixel space. Our teacher generator, on the other hand, is designed to prevent the student generator from collapsing. After network optimization converges, with any given input image, our method will try to generate a 'normal' image that is most similar to the input. **When the input image is similar to the training images (i.e. normal), our framework will generate images in high quality. When the input image differs from any**

of the training data (i.e. abnormal), since our framework was never trained on such cases, the reconstructions will be in low quality and have major differences compared to the inputs. In this way, one can easily alert on anomalies if obvious differences exist by comparing the input and its reconstruction.

Most of existing methods alert on anomaly based on pixel-wise differences with a manually defined threshold. However, as mentioned in §3.4, the noise and distortions exist in the images hamper detection quality. Therefore, we delegate the optimized discriminator network to discriminate anomalies from a perceptual perspective. For notation simplicity, we denote the encoder, teacher generator, student generator, and discriminator as $\mathbf{E}$, $\mathbf{G}_\text{t}$, $\mathbf{G}_\text{s}$, and $\mathbf{D}$ respectively. During inference, $\mathbf{D}$ can be used to compute the anomaly score A via:

$$A = \phi(\frac{\mathbf{D}(\mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I}))) - \mu}{\sigma}), \tag{3.5}$$

where $\phi(\cdot)$ denotes the Sigmoid function, μ and σ are the mean and standard deviation of anomaly scores calculated on all training images.[1]

3.6 Loss Functions

Our SQUID is supervised by a total of five different loss functions. Similar to all existing works, we calculate Mean Square Error (MSE) as the primary self-supervision between input images $\mathbf{I}$ and their reconstructions for both teacher $\mathcal{L}_\text{t}$ and student generators $\mathcal{L}_\text{s}$:

$$\mathcal{L}_\text{t} = (\mathbf{I} - \mathbf{G}_\text{t}(\mathbf{E}(\mathbf{I})))^2 \tag{3.6}$$

and

$$\mathcal{L}_\text{s} = (\mathbf{I} - \mathbf{G}_\text{s}(\mathbf{E}(\mathbf{I})))^2. \tag{3.7}$$

Following the knowledge distillation paradigm, we apply intermediate feature-level supervisions between the teacher and student generators at all generator levels:

$$\mathcal{L}_\text{dist} = \sum_{i=1}^{l} (\mathcal{F}_\text{t}^i - \mathcal{F}_\text{s}^i)^2, \tag{3.8}$$

where l is the total number of generator levels, $\mathcal{F}_\text{t}$ and $\mathcal{F}_\text{s}$ denote the intermediate features in the teacher and student generators, respectively. Additionally, we adopt an adversarial loss [57] to improve the

[1] Early stopping techniques can be adopted for better efficiency.

reconstruction quality and encourage the joint learning of the student generator and the discriminator:

$$\mathcal{L}_{\text{gen}} = \log(1 - \mathbf{D}(\mathbf{G}_{\text{s}}(\mathbf{E}(\mathbf{I})))) \tag{3.9}$$

and

$$\mathcal{L}_{\text{dis}} = \log(\mathbf{D}(\mathbf{I})) + \log(1 - \mathbf{D}(\mathbf{G}_{\text{s}}(\mathbf{E}(\mathbf{I})))). \tag{3.10}$$

By optimizing the adversarial loss, our discriminator seeks to maximize the average of the discrimination scores for real images and the inverted scores for reconstructed ones.

Combining the above five losses together, $\mathbf{E}$, $\mathbf{G}_{\text{t}}$, $\mathbf{G}_{\text{s}}$ with trainable weights θ_{g} are optimized to *minimize* the generative loss:

$$\arg\min_{\theta}(\lambda_{\text{t}}\mathcal{L}_{\text{t}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}}) \tag{3.11}$$

and $\mathbf{D}$ with trainable weights θ_{d} are optimized to *maximize* the discriminative loss:

$$\arg\max_{\theta} \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}, \tag{3.12}$$

where the λs denote the scaling factors that balance each loss term.

3.7 Network Architectures

There are four different networks included in our SQUID: an encoder, a student (main) generator, a teacher generator, and a discriminator. For simplicity and fair comparison, all of the networks are built with the most basic convolutional layers, batch normalization layers, and ReLU activations only.

For an input image of size 128×128, it is first divided into 2×2 non-overlapping patches of size 64×64 each in a pre-processing step. The encoder network is then built upon the architecture parameters as shown in Table 3.1 to extract the patch features.

TABLE 3.1. Encoder architecture 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	$(2 \times 2) \times (4 \times 4)$

As mentioned in §3.1, both the student and teacher generators are built identically. The only difference is that additional learning-based Memory Matrices are placed at different levels of the student generator. The architecture details of the student generator are presented in Table 3.2. Level-wise skip connections from the encoder are only enabled at the levels where Memory Matrices are used. After the last Memory Matrix, the non-overlapping feature patches are put back as a whole for reconstruction.

TABLE 3.2. Student and teacher generator architectures in SQUID. S&M denotes the usage of skip connections and leaning-based 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

Unlike the encoder network and the generator networks, the discriminator is constructed in a more lightweight style. Rather than patch-wise discrimination, reconstructed images are discriminated at their full resolution (i.e. 128×128). The architecture details of the discriminator is presented in Table 3.3.

TABLE 3.3. Discriminator architecture 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

In the tables, the column *Level* denotes the index of feature level, where *Input* denotes the input image and *Output* denotes the reconstruction; the column *#Channels* denotes the number of features, where input images and reconstructions (in gray scale) always have a single feature (pixel) channel; the column *Resolution* denotes the feature map size.

3.8 Creation of DigitAnatomy

Interpreting chest radiographs requires expert knowledge in the medical domain, which may not be friendly for readers from different backgrounds. Considering this, we created a new dataset namely

Algorithm 1 Creation of DigitAnatomy

Input: Pre-loaded MNIST dataset: X, Anomaly types: A = {'normal', 'missing', 'misorder', 'flip', 'zero'}.
Output: Output image: x

```
//Initialize output image to be null
x = ∅
for digit ← 1 to 9 do
    anomaly = random_pick(A)
    //Randomly pick a condition
    if anomaly == 'normal' then
        //Randomly pick a normal digit in the order
        patch = random_pick(X[digit])
    end if
    if anomaly == 'missing' then
        //Anatomy of missing digit
        patch = 0
    end if
    if anomaly == 'misorder' then
        //Anatomy of disorder digit
        misorder_digit = random_pick({1, ··· , 9})
        patch = random_pick(X[misorder_digit])
    end if
    if anomaly == 'flip' then
        //Anatomy of flipped digit
        patch = random_pick(X[digit])
        patch = random_flip(patch)
    end if
    if anomaly == 'zero' then
        //Anatomy of the digit zero
        patch = random_pick(X[0])
    end if
    //Put a patch into the output image
    x[digit] = patch
end for
Reshape x to 2D
```

DigitAnatomy that contains images synthesized by hand written digits collected from the MNIST dataset [39] to verify our main idea better. In this dataset, the human anatomy is translated into the combination of Arabic digits from one to nine in an in-grid placement (see examples in Figure 1.1 and Figure 4.1). We speculate the normal images are the ones containing digits in the correct order; and the types of abnormal images contain the following cases: missing digit, misordered digits, flipped digit(s), and the inclusion of digit zero.

This DigitAnatomy dataset is particularly beneficial for radiography image analysis for the following three reasons: *(i)* It retains two unique properties of radiography images: spatial correlation and consistent shape; *(ii)* Digits are easier for problem shooting and debugging and efforts for interpreting radiography images can be saved; *(iii)* The ground truth of DigitAnatomy is readily accessible, whereas it is hard to collect in radiography images.

The pseudocode for creating DigitAnatomy is shown in Algorithm 1. In practice, we implemented the algorithm as an off-the-shelf data loader that can be integrated into different datasets with few efforts.

CHAPTER 4

Experiments and Results

4.1 Experimental Designs

4.1.1 Public Benchmarks

ZhangLab Chest X-ray [34]. This dataset includes healthy and abnormal pneumonia images, which are officially splitted into training and testing sets. There are 1,349 normal and 3,883 pneumonia images in the training set, and 234 normal and 390 pneumonia images in the testing set. Without accessing the testing data for hyper-parameter tuning, we setup an extra validation set which consists of 200 randomly selected images from the training set with 100 for each of the classes.

Stanford CheXpert [30]. To evaluate our method on larger scale dataset with diverse anomalies, we conducted additional experiments on the front-view PA images in the Stanford CheXpert dataset, which contains a total number of 12 different types of anomalies. In all front-view PA images, the training set is comprised of 5,249 normal and 23,671 abnormal images and the official validation set is comprised of 14 normal and 19 abnormal images. We define a custom testing set consists of 250 normal and 250 abnormal images (with at least 10 images per disease type) picked from the training set.

COVIDx [82]. There are official train/test split in this dataset. The training 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 testing 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.

4.1.2 DigitAnatomy

By implementing Algorithm 1 as an off-the-shelf dataloader, we generate the training data on-the-fly. Similar to the public benchmarks, we setup validation and testing sets separately. A set of 50 images for the normal class and each of the 4 abnormal classes was pre-created, which add up to 250 validation images. For the testing set, the number for each class increases to 200, which add up to 1000.

Without any clinical relevance, calculating the quantitative metrics on this demo dataset is meaningless. Therefore, we mainly present the reconstruction visualizations for more sensible interpretations of SQUID's capability and more intuitive comparisons against the other methods.

4.1.3 Comparing Methods

To demonstrate the superior anomaly detection performance of our proposed method, we consider a total number of **13** primary competing methods for direct comparison. Unless explicitly noted, whenever required, we re-implemented the comparison methods ourselves and trained their models from scratch (disabled pre-training) on our datasets for fair comparisons.

Classic UAD methods. Auto-Encoder and VAE [36]. These classic methods have been widely used for anomaly detection in different domains due to their simplicity. We compare SQUID against them to verify the superiority of our method over the classic ones.

Current UAD state of the arts in medical imaging. Ganomaly [1], f-AnoGAN [64], IF [52], and SALAD [90]. These methods were particularly designed for anomaly detection in medical contents. We compare SQUID against them to verify the superiority in the medical domain.

Recent UAD methods. MemAE [15], CutPaste [43], M-KD [61], PANDA [58], PaDiM [10], and IGD [8]. These methods were published within the recent two years representing the most advanced progress in the general UAD domain. We compare SQUID against them to verify the specialization of our design.

4.1.4 Metrics

In the image-level anomaly detection task, methods' performances are usually measured by the same metrics used for binary classification: Accuracy (Acc), F1-score (F1), Receiver Operating Characteristic (ROC) curve, Area Under the ROC Curve (AUC), and Precision Recall Curve (PRC). Specifically, Acc directly measures how correct the method can detect anomalies regardless of post-processing (e.g. thresholding). This metric is a good reflection for the overall performance when the amount of data is large enough. Similarly, F1-score is also commonly used to reflect on whether the method has balanced performances. Unlike a direct measurement on the portion of correct detections, F1-score is more sensitive to the falsely detected ones. Before calculating F1-score, the confusion matrix with 4 different statistical terms: TP, FP, TN, FN should be measured in advance. TP (True Positive) denotes the method correctly predicts the positive class (i.e.abnormal cases). TN (True Negative) denotes the method correctly predicts the negative class (i.e.normal cases). FP (False Positive) denotes the method model incorrectly predicts the positive class, while the ground truth is actually negative. FN (False Negative) denotes the method incorrectly predicts the negative class, while the ground truth is actually positive. Given the confusion matrix, one can calculate sensitivity and specificity as below:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4.1}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{4.2}$$

Sensitivity measures how many abnormal images are correctly detected and specificity measures how many normal images are correctly classified. Since both measurements are equally important to indicate a method's detection ability, we used F1-score as an unified measurement that combines both sensitivity and specificity through:

$$F1 - score = 2\frac{Sensitivity \cdot Specificity}{Sensitivity + Specificity}. \tag{4.3}$$

Sigmoid gated discrimination scores A are represented as continuous random variables within the range (0,1). Unlike Acc and F1-score that evaluate models' performances based on a given threshold, ROC, AUC, and PRC evaluate models' performances for all possible thresholds. Assume A follows a probability density $f_1(x)$ if the target image x contains anomaly (a positive case) and $f_0(x)$ if otherwise. Then, TPR (True Positive Rate) and FPR (False Positive Rate) are given by:

$$\text{TPR}(T) = \int_T^\infty f_1(x)\,dx, \qquad \text{FPR}(T) = \int_T^\infty f_0(x)\,dx. \tag{4.4}$$

The ROC is then a parametric graph with TPR as the x-axis and FPR the y-axis and AUC denotes the area under such a curve. A method with larger AUC balances better between TPR and FPR.

Unless explicitly specified, we trained all models from scratch for at least *three* times independently, and report both the average metric scores and their standard deviations.

4.1.5 Implementation Details

Pre-processing. Since the datasets contain images of different sizes, it is necessary to resize them into a common size before passing them into the networks. Without losing many visual details, for better computational efficiency, we resized all images to 128×128. We then normalized the intensities of all images into the range [0,1] by using the max and min intensity values collected from the training set.

Data augmentations. Without modifying potential pathological patterns exist in the images, we adopted the most basic augmentations only: *(i)* Random translation within the range $[-0.05\%, +0.05\%]$ in four directions; *(ii)* Random scaling within the range of $[0.95\%, 1.05\%]$.

Network optimization. The Adam [35] optimizer with a batch size of 16 and a weight decay of $1e^{-5}$ was used to optimize our network. The learning rate was initially set to $1e^{-4}$ for both the generator and the discriminator. We considered a learning rate scheduler that automatically adjusts the learning rate following the cosine annealing schedule. The learning rate was eventually decayed to $2e^{-5}$ in a total number of 1,000 epochs. To avoid mode collapse in the generators, we trained them in a period of two iterations while the discriminator, due to its simple architecture, was optimized at every single iteration.

The loss weights were set as $\lambda_{\text{t}} = 0.01$, $\lambda_{\text{s}} = 10$, $\lambda_{\text{dist}} = 0.001$, $\lambda_{\text{gen}} = 0.005$, and $\lambda_{\text{dis}} = 0.005$. Input images were divided into 2×2 non-overlapping patches. The probability for masking shortcut was set to $\rho = 95\%$. We set $k = 5$ most similar features to be activated in the Gumbel Shrinkage. These parameter choices were particularly studied in §4.4. For better consistency, the same set of hyper-parameters found in the ZhangLab experiments was also used for experiments on all other datasets.

Post-processing. After obtaining the discrimination scores, we followed the normalization schema as mentioned in Equation 3.5 to normalize them based on the training set statistics. We then used the validation set to decide the best threshold for deterministic predictions on the testing set.

4.2 Image Reconstruction Results on DigitAnatomy

Figure 4.1 presents qualitative results on DigitAnatomy to examine the capability of SQUID and to interpret the mistakes made by existing methods [1, 15, 64]. We deliberately inject anomalies (e.g. misordered, missing digits) into normal images (highlighted in red) and test if the model can reconstruct their normal correspondences. We also assess the reconstruction quality from a blank image (as an extreme case) to raise the task difficulty. 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), our in-painting block will augment the abnormal feature to its normal correspondence by assembling top-k most similar patterns from the dictionary. Other methods, however, do not possess this ability, so they reconstruct defective images. For example, GAN-based methods (f-AnoGAN and Ganomly) tend to reconstruct an exemplar image averaged from all training examples. MemAE performs relatively better than f-AnoGAN and Ganomly due to the usage of a memory module. It does not work well for the anomaly of missing digits and completely fails on the extreme anomaly attack.

4.3 Results on Public Datasets

Our SQUID was evaluated on three publicly available datasets as mentioned earlier: ZhangLab Chest X-ray, Stanford CheXpert, and COVIDx to compare against a wide range of state-of-the-art counterparts.

On the ZhangLab Chest X-ray dataset, as reported in Table 4.1, SQUID achieves the best detection result on all of the metrics. Specifically, SQUID outperforms the second best runner-up counterpart SALAD [90] by 4.9% in AUC, 6.7% in Acc, and 2.6% in F1. SQUID yields even higher performance improvements when comparing to the methods published in recent 2 years.

Our method also achieves the best results on all of the metrics for the Stanford CheXpert dataset as shown in Table 4.2. With 12 different types of diseases, the Stanford CheXpert dataset is much more challenging than the ZhangLab dataset: all methods underwent an obvious performance drop. Even in this way, SQUID still achieves an average AUC of 78.1%, Accuracy of 71.9%, and F1 of 75.9% that are 8.3%, 5.9%, and 12.3% higher than the second best runner-up method M-KD [62].

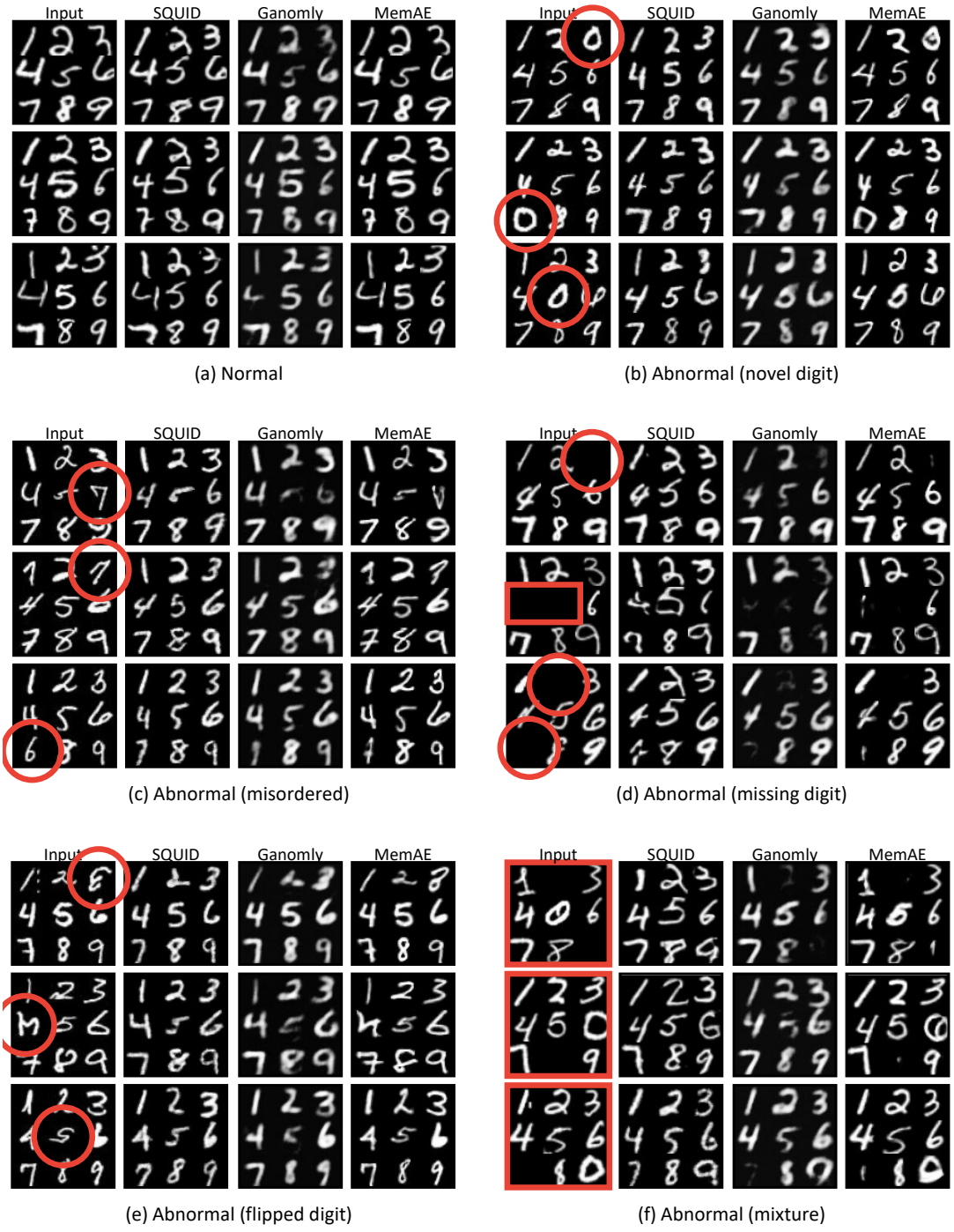


FIGURE 4.1. **Reconstruction results on DigitAnatomy.** Our feature-level in-painting approach is more robust to amplified noise and pixel variance than the existing pixel-level in-painting methods. Major anomalies are highlighted in red.

TABLE 4.1. Results on the test sets of the ZhangLab dataset. Both average results and standard deviations are reported. † denotes the results taken from other literature.

ZhangLab	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Auto-Encoder†	-	59.9	63.4	77.2
VAE† [36]	Arxiv'13	61.8	64.0	77.4
Ganomaly† [1]	ACCV'18	78.0	70.0	79.0
f-AnoGAN† [64]	MIA'19	75.5	74.0	81.0
MemAE [15]	ICCV'19	77.8±1.4	56.5±1.1	82.6±0.9
MNAD [55]	CVPR'20	77.3±0.9	73.6±0.7	79.3±1.1
SALAD† [90]	TMI'21	82.7±0.8	75.9±0.9	82.1±0.3
CutPaste [43]	CVPR'21	73.6±3.9	64.0±6.5	72.3±8.9
PANDA [58]	CVPR'21	65.7±1.3	65.4±1.9	66.3±1.2
M-KD [62]	CVPR'21	74.1±2.6	69.1±0.2	62.3±8.4
IF 2D [52]	MICCAI'21	81.0±2.8	76.4±0.2	82.2±2.7
PaDiM [10]	ICPR'21	71.4±3.4	72.9±2.4	80.7±1.2
IGD [8]	AAAI'22	73.4±1.9	74.0±2.2	80.9±1.3
SQUID	This work	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**

TABLE 4.2. Results on the test sets of the CheXpert dataset. Both average results and standard deviations are reported.

CheXpert	Ref & Year	AUC (%)	Acc (%)	F1 (%)
Ganomaly [1]	ACCV'18	68.9±1.4	65.7±0.2	65.1±1.9
f-AnoGAN [64]	MIA'19	65.8±3.3	63.7±1.8	59.4±3.8
MemAE [15]	ICCV'19	54.3±4.0	55.6±1.4	53.3±7.0
CutPaste [43]	CVPR'21	65.5±2.2	62.7±2.0	60.3±4.6
PANDA [58]	CVPR'21	68.6±0.9	66.4±2.8	65.3±1.5
M-KD [62]	CVPR'21	69.8±1.6	66.0±2.5	63.6±5.7
SQUID	This work	**78.1±5.1**	**71.9±3.8**	**75.9±5.7**

TABLE 4.3. Results on the test sets of the COVIDx dataset. Both average results and standard deviations are reported. † denotes the results are taken from [69]. ‡ denotes the results are taken from [78].

COVIDx	Ref & Year	AUC (%)	Acc (%)	F1 (%)
PaDiM† [10]	ICPR'21	54.0	-	-
Ganomaly† [1]	ACCV'18	58.4	-	-
f-AnoGAN‡ [64]	MIA'19	66.9	-	-
MemAE [15]	ICCV'19	71.8±3.6	77.1±2.1	86.4±0.8
PANDA [58]	CVPR'21	72.3±1.0	**76.9±0.8**	**86.4±0.4**
M-KD [62]	CVPR'21	71.7±1.1	69.7±4.5	55.6±2.5
SQUID	This work	**74.7±0.9**	76.8±0.1	86.0±0.2

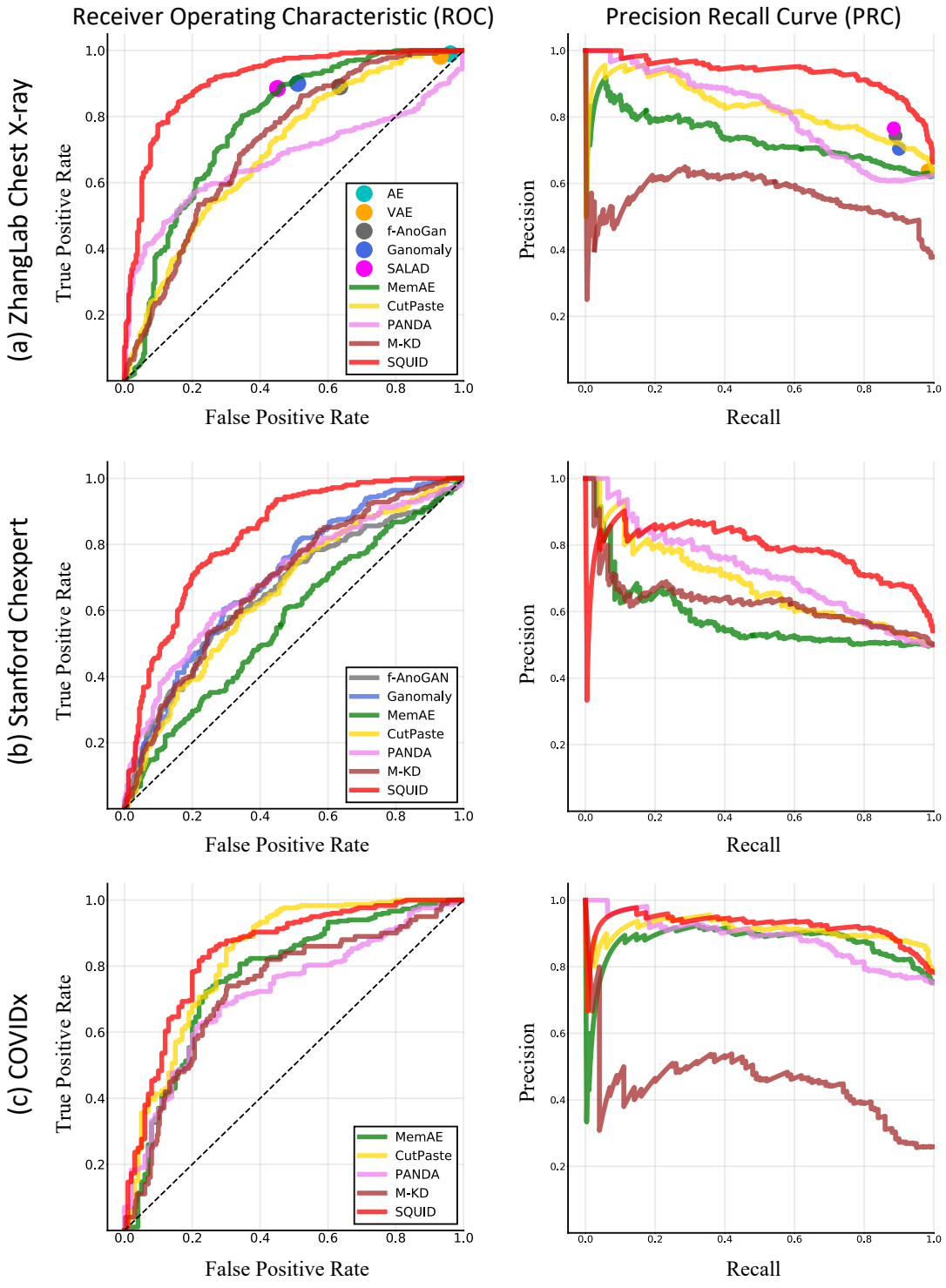


FIGURE 4.2. **ROC and PRC comparison.** SQUID yields the best ROC and PRC in all of the comparison methods for all 3 datasets.

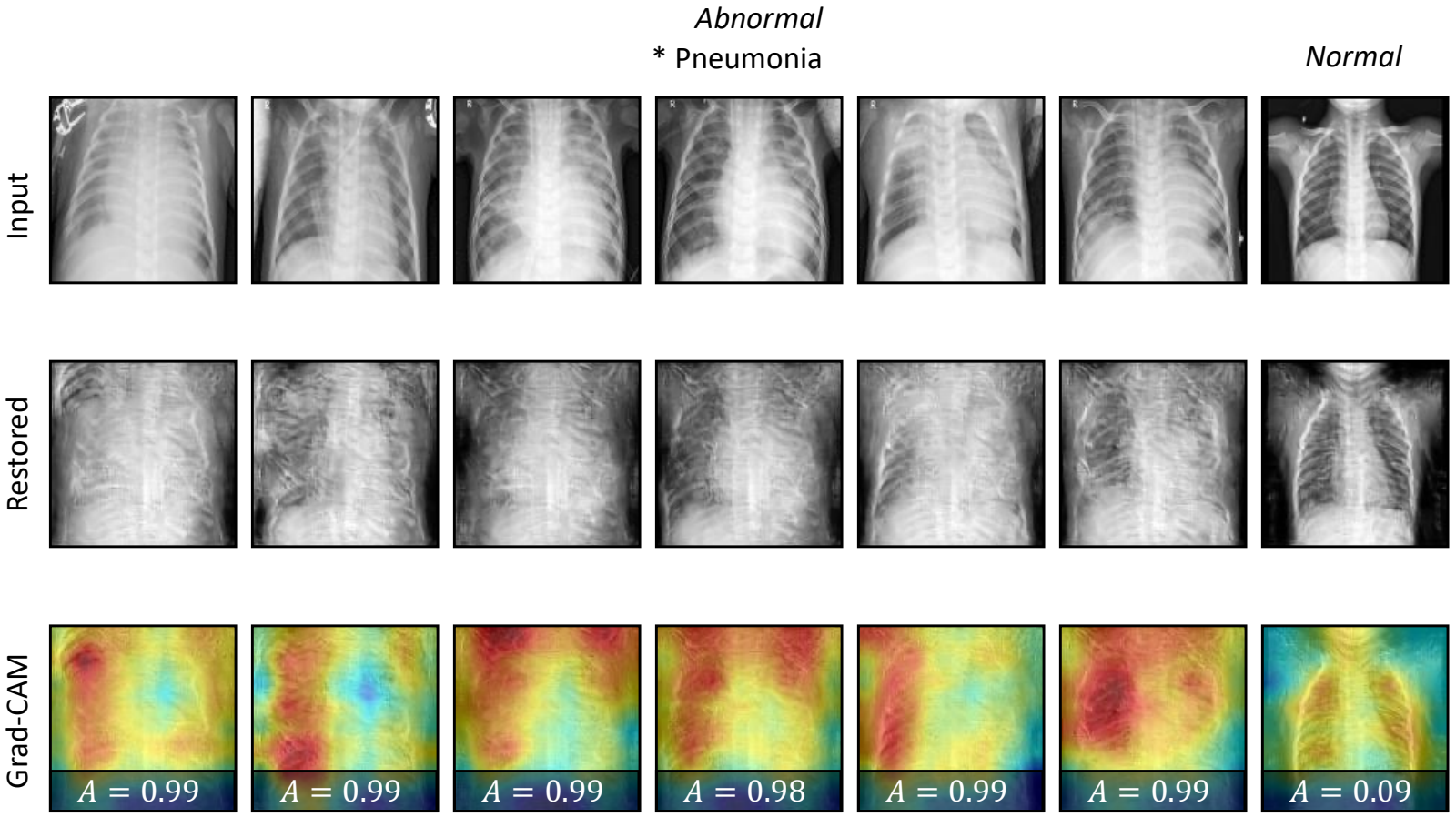


FIGURE 4.3. **Reconstruction results of SQUID on the ZhangLab Chest X-ray dataset.** Normal and abnormal (Pneumonia) cases are separated in different rows. Corresponding Grad-CAM heatmaps along with anomaly scores are shown as well.

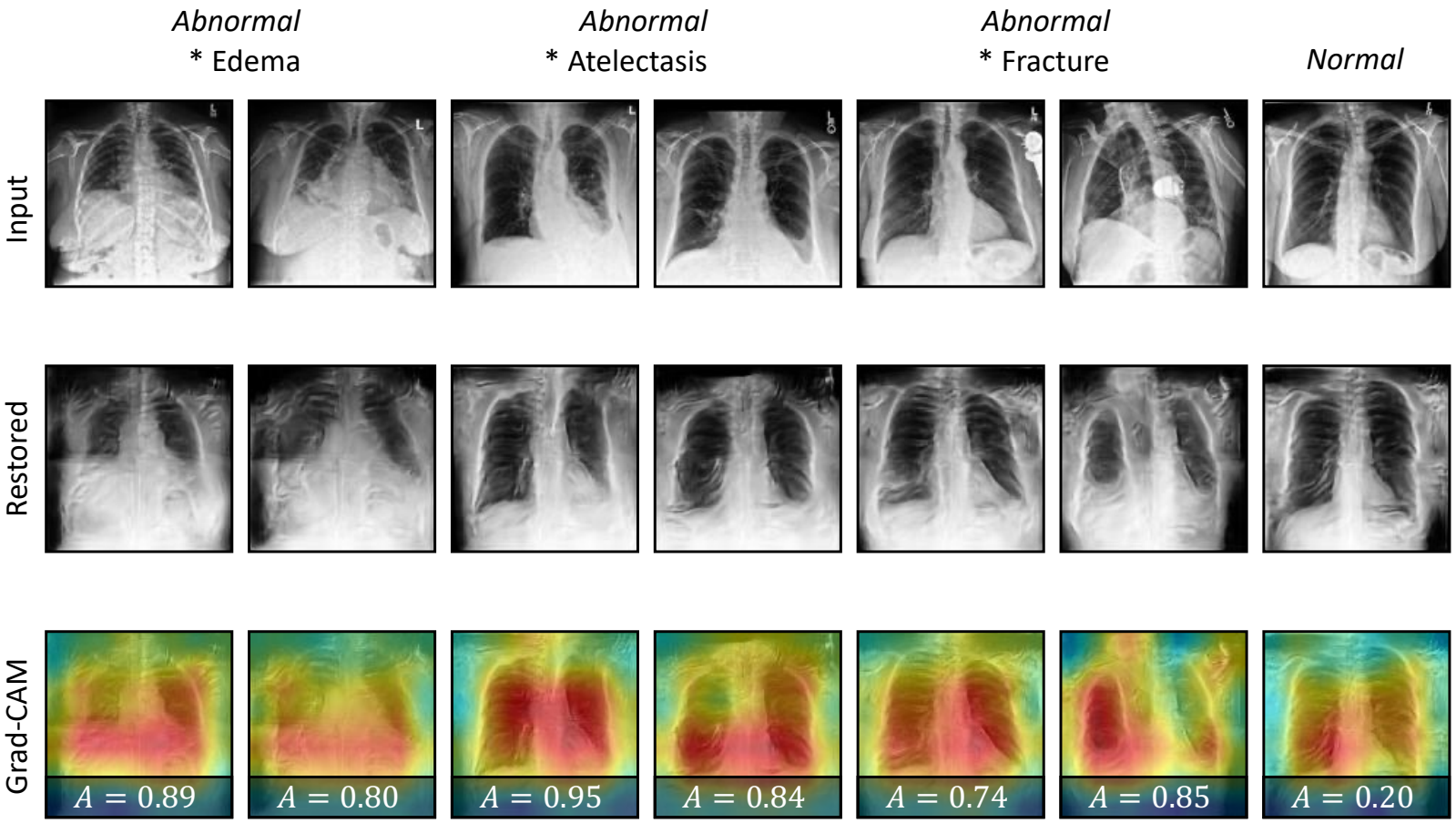


FIGURE 4.4. **Reconstruction results of SQUID on the Stanford CheXpert dataset.** Diseases including: Fracture, Atelectasis, and Edema are separated in different rows. Corresponding Grad-CAM heatmaps along with anomaly scores are shown as well.

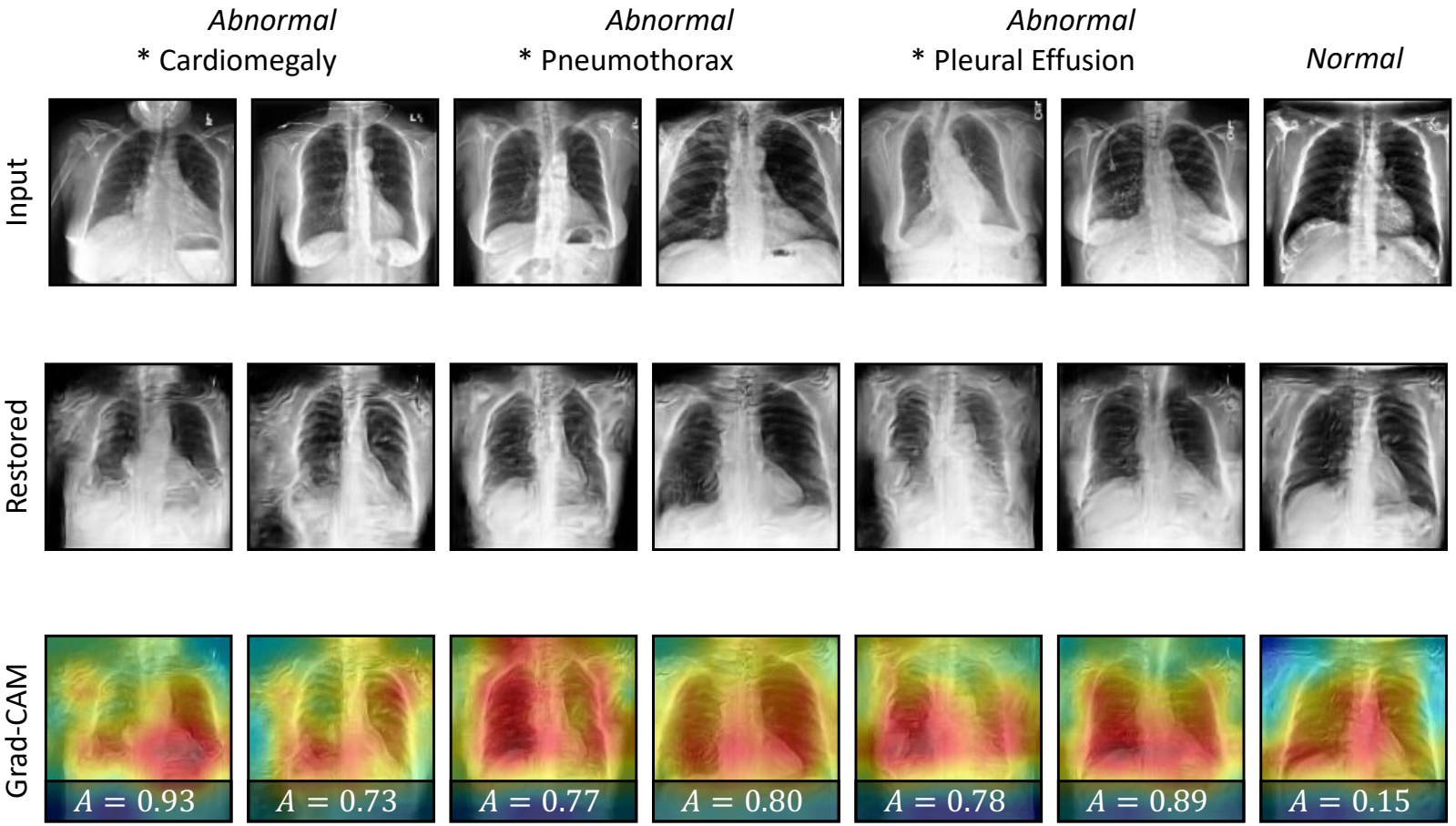


FIGURE 4.5. **Reconstruction results of SQUID on the Stanford CheXpert dataset.** Diseases including: Pleural Effusion, Pneumothorax, and Cardiomegaly are separated in different rows. Corresponding Grad-CAM heatmaps along with anomaly scores are shown as well.

Unlike common pneumonia, Covid infections are more concealed and much harder to be detected in chest X-rays. To further explore the limit of our method on real-world applications, we conducted comparison experiments on the COVIDx dataset and report Covid infection detection results in Table 4.3. As reflected in the table, all advanced deep learning algorithms demonstrate limited results on this challenging dataset. However, SQUID stands out with its superior average AUC score of 74.7% that is 3.0% higher than M-KD [62] and 2.9% higher than MemAE [15].

The Receiver Operating Characteristic curve (ROC) and Precision Recall Curve (PRC) on the three datasets are presented in Figure 4.2. Our method yields the best trade-off between sensitivity and specificity. Overall, the significant improvements observed on SQUID proved its effectiveness.

In Figure 4.3, Figure 4.4, and Figure 4.5 we visualize the reconstruction results of SQUID on exemplary normal and abnormal images in the ZhangLab Chest X-ray and the Stanford CheXpert datasets. For normal cases, SQUID can easily find a similar match in the memory and achieve the reconstruction smoothly. For abnormal cases, contradictions will arise when imposing forged normal patterns into the abnormal features. In this way, the generated images will vary significantly from the input, which can be captured by the discriminator. We plot the heatmap of the discriminator (using Grad-CAM [68]) to indicate the regions that are most likely to appear anomalous. As a result, the reconstructed healthy images yield much lower anomaly scores than the diseased ones, validating the effectiveness of SQUID.

4.4 Ablation Studies

Component study. We first examine the impact of every proposed component in SQUID by taking each one of them out of the complete framework. Table 4.4 shows that each component accounts for at least 5% performance gains. The space-aware memory (+10.0%) and in-painting block (+6.7%) are among the most significant contributors, which underline our motivation and justification of the method development (§3.2 and §3.4). 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 [31], stop gradient [23], and masked shortcut [85]) are necessary for the remarkable performance. Although replacing Memory Queue with Memory Matrix could still maintain a decent

result (only dropped by 5.1%), our Memory Queue presents a more trustworthy recovery of "normal" patterns in the image than the baseline Memory Matrix [15] evidenced by Figure 4.1.

TABLE 4.4. Component studies indicate that the overall performance benefits from all of the components in SQUID. The ablation study is conducted on the ZhangLab dataset.

Method	AUC(%)	Acc(%)	F1(%)
w/o Space-aware Memory	77.6±0.5	75.5±0.5	82.5±0.6
w/o In-painting Block	80.9±2.1	75.8±1.5	81.6±1.3
w/o Skip Connection	79.5±1.6	73.0±1.4	78.8±0.5
w/o Hierarchical Memory	82.9±1.2	77.4±1.1	81.2±0.5
w/o Knowledge Distillation	85.4±0.8	79.5±0.7	83.5±0.8
w/o Stop Gradient	85.0±4.3	77.6±2.8	79.8±1.6
w/o Gumbel Shrinkage	86.2±3.3	80.5±3.2	85.4±2.1
Full SQUID	**87.6±1.5**	**80.3±1.3**	**84.7±0.8**

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 Matrix are three important hyper-parameters of SQUID. Here, we conducted exhausted experiments on these parameters in Figure 4.6. Trials were first made on the number of patches from 1×1 to 8×8. When dividing input images into a single patch, space-aware settings are not triggered, hence yielding the worst performance. Although the spatial structures are relatively stable in most chest X-rays, certain deviations can still be observed. Therefore, with small patches, object parts in one patch can easily appear in adjacent patches and be misdetected as anomalies. The number of topk activations in Gumbel softmax also impacts the performances. By assembling the top-5 most similar patterns through Gumbel softmax, SQUID is able to achieve the best result. When replacing input features with the top-1 most similar pattern, SQUID suffers from a performance drop by -15% 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 200 items per region. When the item number exceeds 500 per region, AUC scores begin to drop continuously.

4.5 Extensive Studies

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

Convolution vs. transformer Layers. In our proposed in-painting block, a transformer layer is used to aggregate the encoded patch features and the Memory Queue augmented "normal" features. However,

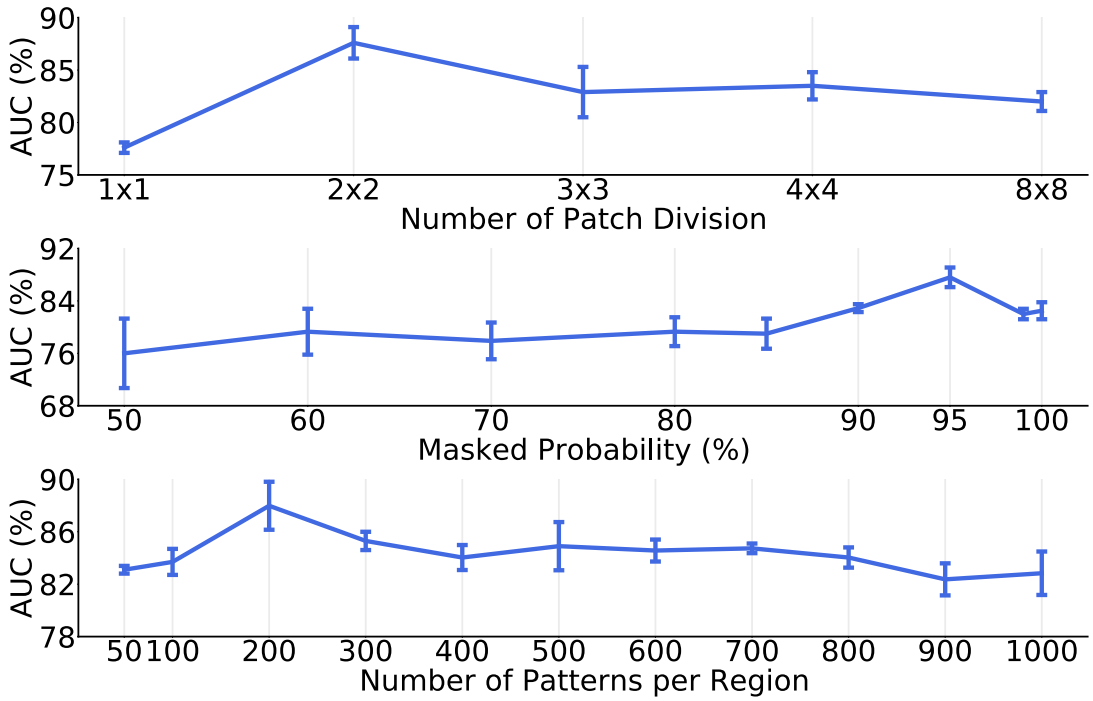


FIGURE 4.6. **SQUID is robust to hyper-parameter modifications.** The best result is obtained at dividing 2×2 patches, setting 200 patterns per memory region, and activating top 5 patterns through Gumbel Shrinkage.

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.

Soft vs. hard masked shortcuts. In our proposed masked shortcut, skipped and in-painted features are aggregated using a binary gating mask. An 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. 3.3, we conducted experiments by aggregating the patch features $\mathcal{F}$ through:

$$\mathcal{F}' = (1 - \rho) \cdot \mathcal{F} + \rho \cdot \texttt{inpaint}(\mathcal{F}), \tag{4.5}$$

where ρ was set to 95%, following the best setting adopted.

Pixel-level vs. feature-level in-painting. As discussed in §3.4, 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 [44, 56, 92]. To validate our claim, we 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,

TABLE 4.5. 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**

we randomly zeroed out parts of the input patches with 25% probability and let SQUID in-paint the masked input images. All other settings and objective functions remain unchanged.

Effectiveness of our space-aware settings. MemAE [15] 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 their unique space-specific patches and were then evaluated through an ensemble style to compare with our SQUID.

TABLE 4.6. To test the effectiveness of our space-aware settings, we apply them to MemAE [15]. 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 [15]	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**

Summary. The results of the first three ablative experiments are presented in Table 4.5. 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.

The results for space-aware MemAE are presented in Table 4.6. 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 encoded this

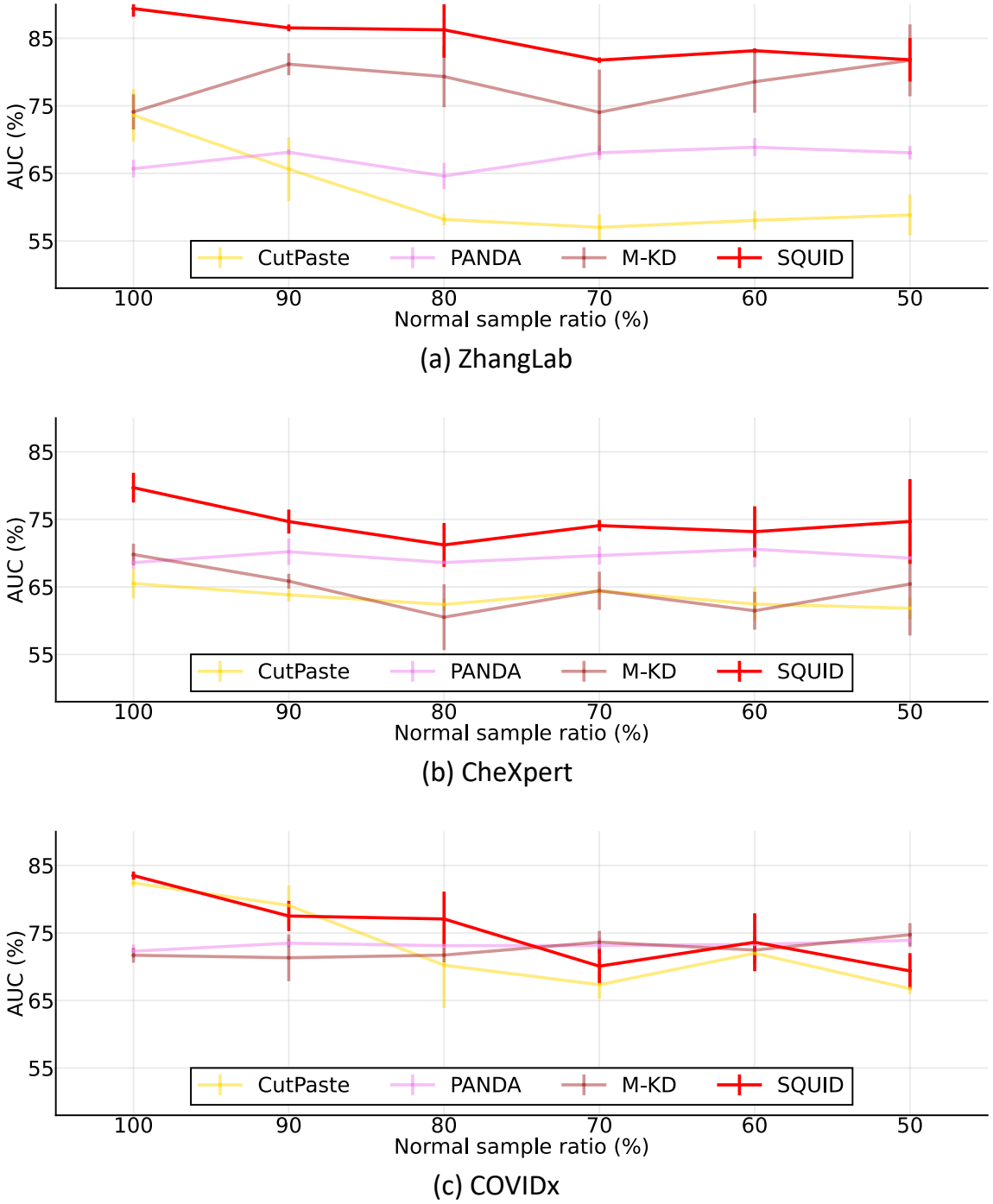


FIGURE 4.7. **Ablation study of mixing normal and abnormal samples in the training set.** SQUID is robust to mixed training with different normal/abnormal ratios on the ZhangLab, CheXpert, and COVIDx datasets.

spatial information into the feature dictionary, ultimately requiring only one model. Both effectiveness and efficiency are pronounced.

4.6 Robustness to Disease Samples in the Training Set

In the main stream training protocol, unsupervised anomaly detection methods are trained on the dataset contains normal data only. However, collecting such pure normal datasets requires implicit weak supervision. Note that there is no work investigating the robustness of "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 SQUID with three competitive baselines (CutPaste [43], PANDA [58], and M-KD [61]) that originally relies on a pure normal training set. Figure 4.7 remarks that SQUID is robust to the disease/healthy training ratio up to 50% and maintains AUC above 0.8 on the ZhangLab dataset by automatically omitting minority anatomical patterns. On the Stanford CheXpert and the COVIDx dataset, SQUID still achieves better or comparable AUC to abnormal training samples than the baseline models. In contrast, CutPaste drops significantly as the percentage of disease-free images decreases; PANDA and M-KD can maintain good performances due to their advanced framework design. Interestingly, M-KD with mixed training data even outperforms its original training setting, although considerable fluctuations can be observed.

CHAPTER 5

Discussion

5.1 Limitations

Evidenced by the superior experimental results, we believe SQUID is able to detect anomalies in chest X-rays under challenging scenarios and it is adequate to be deployed in open source software for real-world applications. However, we found SQUID, in its current form, still has unneglectable limitations that should be discussed and resolved in the future.

Complex framework design. One of the biggest limitations of our method is its complexity. SQUID consists of three sequential stages and four individual networks. Compared to the most simple methods (e.g. Auto-Encoder and VAE) that usually demand only two simple networks, our method yields exceeding number of parameters and huge computational overhead. It is also a time consuming process to train our networks: every single training image needs to be processed by the four networks that consumes additional hardware memory inevitably.

Inefficient framework inference. After optimizing the framework, it is also important to measure the overall inference time when detecting on unseen images. This becomes more critical when real time X-ray detection (e.g. $< 30ms/image$) is desired. In an ideal supervised setting with image-wise ground truths, a single encoder-style network can be trained just like binary classification. However, such efficient inference is impossible in the unsupervised setting. SQUID, same to other advanced methods (e.g. Ganomaly, SALAD etc.), detects anomalies in an image through the forward pass of three networks—encoder→generator→discriminator.

Inaccurate pixel-wise anomaly detection. SQUID is designed to provide as accurate diagnosis as possible for appearances of abnormal conditions. However, in rare scenarios, where users are more interested for finer-gradient diagnosis with pixel-wise disease masks, SQUID is not adequate in the

current form. It is understandable because in the unsupervised setting, pixel-wise anomaly localization is much harder than image-wise detection. Although one can calculate Grad-CAM masks (as in Figure 4.3, Figure 4.4, and Figure 4.5) as coarse indications of potential disease areas, they are not precise enough to be regarded as pixel-wise localization masks. Note that there are indeed other methods that were particularly proposed toward pixel-wise anomaly detection [61, 73, 77, 83, 89]. Most of those methods compute the localization mask as pixel-level residuals between input and their reconstruction. However, such an approach suffers from input distortions and can hardly generalize to images collected from different data sources.

5.2 Future Work

Despite the above limitations, SQUID demonstrates great potentials to be further extended in future research. Anticipated future work should be designed to improve either SQUID's efficiency, detection accuracy, or better pixel-wise localization ability.

Towards better efficiency. The most straightforward strategy for efficiency improvement is to slim the networks. One can simply reduce the feature channels for every intermediate layers to obtain direct inference speed ups and smaller model size. However, as a trade-off, overall detection accuracy will drop accordingly due to the loss of representation ability.

Instead of directly shrinking feature channels, another possible strategy is to adopt advanced pruning and quantization strategies [21] to spot sparser sub-architectures to function in a low-precision environment.

Replacing the basic convolution layers with more advanced operators such as separable convolutions [28], ghost convolutions [20], micro convolutions [45] etc. is also helpful. These operators not only reduce computations but also maintain network's performance to the greatest extent. Since these operators were validated on common vision tasks for natural contents only, it is essential to re-evaluate their capabilities on this particular task domain.

Inspired by a recent work [86], skip connections used between our encoder network and generators can be further upgraded into bi-directional skip connections. By iterating the encoded and decoded features through such skip connections, network layers can be reused. Therefore, without crafting a large network with exceeding number of layers and feature channels, bi-directional skip connections help us to construct SQUID in a light-weight style.

It is also a good research direction to dig into the memory module further. In our Space-aware Memory, images are first segmented into patches and then being searched within non-overlapping memory regions. Concurrent to spatial dimensions, one can also explore Channel-aware Memory to perform search individually with respect to different feature channels.

Towards better anomaly detection accuracy. As ablated in §4.4, our method could benefit from a careful design of the network components. We believe by enhancing the feature representation ability of the backbone network, improvements on down-streaming tasks (e.g. anomaly detection) are also accessible. One can obtain this enhancement easily by upgrading our network architectures to more advanced stat-of-the-art backbone networks (e.g. ConvNext [49], ViT [12] etc.). Still, as a trade-off, large networks usually lead to worse computational efficiency.

Without modifying backbone networks, improvements on detection accuracy can also be achieved by proper data augmentations [43, 90]. We note a recent work [63] approaches chest X-ray anomaly detection by simulating the appearances of all possible anomalies. The authors have conducted thorough empirical studies to mainly compare against our SQUIDand their results reflect a slight performance improvement. As a future direction, one can build a hybrid framework includes both advanced augmentation techniques and the proposed in-painting strategy.

The experimental results reported in Chapter 4 indicate that our method could benefit from the usage of large datasets as well. In this way, more training features could be memorized in the network and SQUID could provide better feature discriminations when inferring on unseen data. This observation reminds us of generative models. Given the available dataset, one can 'forge' similar data to fit into the distribution of the given dataset by using Generative Adversarial Nets (GANs) [17] or Denoising Diffusion Probabilistic Models (DDPMs) [27].

Towards pixel-wise localization. A naive extensive from SQUID to enable pixel-wise localization is to calculate pixel-wise differences between input images and the reconstructed ones. The calculated residuals can be then binarized to a localization mask with a certain threshold. However, as discussed in Chapter 2, comparing images in the raw pixel space can be problematic due to unexpected corruptions (e.g. noise, artifacts). Considering this, our feature space in-painting makes it handy to calculate the residuals between their encoded semantics. This can be achieved by feeding the reconstructed images back into the encoder again.

In recent studies, contrastive learning approaches [6] with vision transformers [12] are capable of learning attentive regions in a self-supervised learning manner. This kind of trait can be naturally incorporated into unsupervised anomaly detection. The attention networks can automatically highlight regions of interests that are responsible for the final detection.

CHAPTER 6

Conclusion

In this work, we present SQUID for unsupervised anomaly detection in radiography images. Existing methods either use neural networks to 'memorize' feature patterns of normal data or manually simulate fake anomalies through data augmentations. Although both kinds of methods demonstrated good anomaly detection performances, those methods are limited under certain circumstances: Memorizing training data is impractical for training on normal-abnormal mixtures and data augmentations cannot simulate all possible types of anomalies.

The success of memory networks inspires us to design more robust and efficient memory modules to generalize normal patterns. Unlike traditional memory network methods that fit the encoded features of the input image into the memory matrix as a whole, we propose the Space-aware Memory to function on non-overlapping image patches separately. Such an update lowers the overall complexity and is especially suitable for chest X-ray scans with consistent anatomical patterns. Additionally, for better fitting into the distribution of training data, we disable memory matrix learning and directly copy training features into the matrix by following a queue-based protocol.

To avoid mode collapse caused by the Memory Matrix, we adopt the knowledge distillation and the gradient-stopping techniques to build an extra teacher network as a regularizer. After reconstructing input images into their most similar normal images, we utilize another discriminator network to asses the quality of the reconstructed images and eventually alert on the poorly reconstructed ones.

In experiments, SQUID demonstrated the best anomaly detection performances over all existing state-of-the-art competing methods. Qualitatively, we show that SQUID can taxonomize the anatomical structures into exemplary patterns; and our novel framework design indeed identifies unseen/modified patterns in the images (supported by Grad-CAM heatmaps). Quantitatively, SQUID yields the best detection results by exceeding the second best runner-up method over 5% AUC on the ZhangLab dataset; 10% AUC improvement on the Stanford CheXpert dataset; and 3% AUC improvement on the COVIDx

dataset. The superior quantitative results again validated our observations: *Radiography imaging protocols function on particular body regions, therefore generating images of great similarity and yielding repeated anatomical structures across patients.*

For better explainability, we further created a novel dataset, namely DigitAnatomy that consists of combinations of hand-written digits only. We believe this new dataset is able to synthesize the spatial correlation and consistent shape of chest anatomy in radiography images. It can be utilized to prompt the development, evaluation, and interpretability of anomaly detection methods, particularly in the radiography imaging community.

Bibliography

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

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

[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. 2016. Automatic segmentation of wrist bones in ct using a statistical wrist shape + pose model. *IEEE transactions on medical imaging*, 35(8):1789–1801.

[4] Arpan Basu, Riktim Mondal, Showmik Bhowmik, and Ram Sarkar. 2020. U-net versus pix2pix: a comparative study on degraded document image binarization. *J. Electronic Imaging*, 29(6):063019.

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

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

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

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

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

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

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

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image

recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

[14] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. 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.

[15] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714.

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

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

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

[20] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. 2020. Ghostnet: More features from cheap operations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1577–1586. Computer Vision Foundation / IEEE.

[21] Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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

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

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

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

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

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[28] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

[29] Yu Huang and Yue Chen. 2020. Autonomous driving with deep learning: A survey of state-of-art technologies. *CoRR*, abs/2006.06091.

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

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

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

[33] Tero Karras, Samuli Laine, and Timo Aila. 2021. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228.

[34] 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. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.

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

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

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.

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

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324.

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

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

[42] Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee Kim. 2018. 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.

[43] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. 2021. 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.

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

[45] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Lu Yuan, Zicheng Liu, Lei Zhang, and Nuno Vasconcelos. 2021. Micronet: Improving image recognition with extremely low flops. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 458–467. IEEE.

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

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

[48] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. 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.

[49] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

[51] Zahra Mirikharaji and Ghassan Hamarneh. 2018. 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.

[52] Sergio Naval Marimont and Giacomo Tarroni. 2021. Implicit field learning for unsupervised anomaly detection in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–198. Springer.

[53] Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. 2021. 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.

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

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

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

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

[58] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. 2021. 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.

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

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

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

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

[63] Junya Sato, Yuki Suzuki, Tomohiro Wataya, Daiki Nishigaki, Kosuke Kita, Kazuki Yamagata, Noriyuki Tomiyama, and Shoji Kido. 2022. Anatomy-aware self-supervised learning for anomaly detection in chest radiographs. *CoRR*, abs/2205.04282.

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

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

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

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

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

[69] Md Mahfuzur Rahman Siddiquee, Teresa Wu, and Baoxin Li. 2021. A2b-gan: Utilizing unannotated anomalous images for anomaly detection in medical image analysis.

[70] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. 2019. 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.

[71] 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. 2017. 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.

[72] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[73] Krishna Kumar Singh and Yong Jae Lee. 2017. 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.

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

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

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

[77] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. 2018. 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.

[78] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan Verjans, Rajvinder Singh, and Gustavo Carneiro. 2021. Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images.

[79] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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

[81] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. 2021. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364.

[82] Linda Wang, Zhong Qiu Lin, and Alexander Wong. 2020. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12.

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

[84] Tiange Xiang, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, and Weidong Cai. 2020. Bio-net: Learning recurrent bi-directional connections for encoder-decoder architecture. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*, volume 12261 of *Lecture Notes in Computer Science*, pages 74–84. Springer.

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

[86] Tiange Xiang, Chaoyi Zhang, Xinyi Wang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. 2022. Towards bi-directional skip connections in encoder-decoder architectures and beyond. *Medical Image Anal.*, 78:102420.

[87] Muhammad Zaigham Zaheer, Arif Mahmood, M Haris Khan, Marcella Astrid, and Seung-Ik Lee. 2021. 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.

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

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

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

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

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

[93] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. 2019. 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.

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

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