Deep Semi-supervised Anomaly Detection Using VQ-VAE

Renuka Sharma[1,2,3,4], Hengcan Shi[1] Jianfei Cai[1], Suyash P. Awate[2], Nick Birbilis[5]
[1]Monash University, Melbourne, Australia, [2]Indian Institute of Technology, Mumbai, India,
[3]IITB-Monash Research Academy, Mumbai, India, [4]CSIRO's Data61, Brisbane, Australia
[5]Deakin University, Melbourne, Australia
renuka.sharma@data61.csiro.au, hengcan.shi@monash.edu, jianfei.cai@monash.edu,
suyash@cse.iitb.ac.in, nick.birbilis@deakin.edu.au

***Abstract*—Anomaly detection is a fundamental and challenging task in computer vision, which determines whether an image contains anomaly or not. Prior works using autoencoders for anomaly detection are based on pixel-wise learning in the continuous latent space, which is inefficient since images contain a lot of redundant information. Meanwhile, for most of the anomaly detection methods, the training set only contains normal data due to the unavailability or paucity of labeled anomalous data. However, an exposure to a fraction of labeled anomalous images, even infinitesimal in size in comparison to the amount of normal data, can significantly improve the anomaly detection performance while slightly increasing labeling costs. In this paper, we propose a Semi-Supervised Vector Quantized Variational Autoencoder (ss-VQ-VAE) for anomaly detection. Our ss-VQ-VAE leverages discretized latent space embeddings of VQ-VAE [1] to reduce noise and redundancies for better reconstruction of normal data in comparison with anomalous data. At the core of ss-VQ-VAE, we introduce a new loss to incorporate a few anomalous images available to train the model. In addition, based on the VQ-VAE architecture, we further propose an anomaly score that compares the encoded features of the input with the dictionary embeddings in VQ-VAE to make more accurate predictions. Experimental results on two datasets, MVTec and the corrosion dataset, show the significance of the novelties in our method. The code is available online[1].**

***Index Terms*—Semi-supervision, Vector Quantization, Anomaly Detection, One-class Classification.**

I. Introduction

Anomaly detection is a one-class classification problem which determines whether an image contains anomaly or not. It serves as a crucial step in many higher-level applications such as error and corrosion detection in manufacturing industries [2]–[4], disease detection in medical domain [5], road anomaly detection for automatic driving [6]–[8], anomaly detection in hyperspectral imagery [9] as well as fraud detection in security domain [10], [11].

Prior works often use autoencoders [12]–[15], with and without variational learnings (VAEs), for anomaly detection, which rely on fine continuous latent space embeddings for the normal data points lying on a high dimensional manifold. However, their continuous latent space embeddings often contain noise and information redundancies, which might result in inaccurate image reconstruction and hence incorrect anomaly prediction in the presence of only reconstruction based losses in their learning objectives.

Meanwhile, a straightforward way to train an anomaly detection model is to employ fully-supervised learning strategies with both normal and anomalous training data. However, acquiring such data is time-consuming and labor-intensive. Therefore, previous methods [12], [16] often tackled anomaly detection as an unsupervised problem, where only normal images are in the training set while expecting to classify an image as normal or anomalous during inference. Though anomalies vary in shape and size across categories and are unseen during training, anomaly detection results can still be significantly improved when the training set is exposed with a fraction of anomalous images, as shown in DeepSAD [13].

Based on the above analyses, we propose ss-VQ-VAE (Semi-supervised Vector Quantized Variational Autoencoder), an end-to-end one-class classification model for anomaly detection in this paper. Our ss-VQ-VAE leverages the VQ-VAE [1] model where the latent space embedding is discrete rather than continuous to generate high-fidelity reconstruction. Moreover, we design semi-supervised losses for the VQ-VAE model in anomaly detection, which fuse normal and a fraction of anomalous data to learn a more discriminative dictionary in ss-VQ-VAE. We further introduce an anomaly score to better adapt VQ-VAE to anomaly detection, which compares the encoded features of the input with the dictionary embeddings. More accurate anomaly detection results can be obtained by our anomaly score. Experimental results on the public MVTec dataset show the effectiveness of our method. In addition, we also curate an industrial inspection corrosion dataset for semi-supervised anomaly detection and show our superior performance and applicability on it.

II. Related Work

Early approaches for anomaly detection used hand-crafted features for information extraction. With the advent of larger datasets, it has become difficult to hand-craft the features for a dataset. The kernel-based methods like OC-SVM [17], SVDD [18] and variants of KPCA [19], [20] (Kernel Principal

[1] https://github.com/RenukaSharma/ss-vq-vae

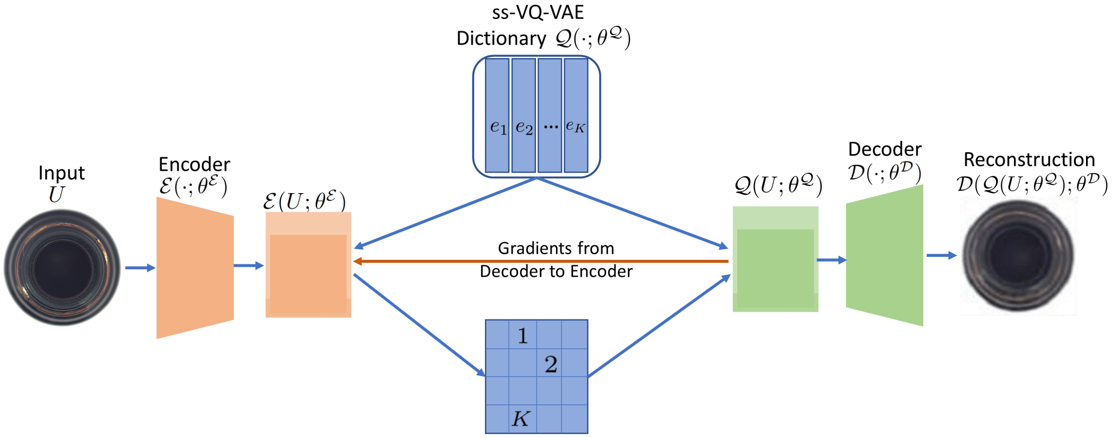

ss-VQ-VAE
Dictionary $\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$
e_1
e_2
···
e_K
Input
U
Encoder
$\mathcal{E}(\cdot;\theta^{\mathcal{E}})$
$\mathcal{E}(U;\theta^{\mathcal{E}})$
Gradients from
Decoder to Encoder
1
2
K
$\mathcal{Q}(U;\theta^{\mathcal{Q}})$
Decoder
$\mathcal{D}(\cdot;\theta^{\mathcal{D}})$
Reconstruction
$\mathcal{D}(\mathcal{Q}(U;\theta^{\mathcal{Q}});\theta^{\mathcal{D}})$

Fig. 1. **Semi-supervised VQ-VAE learning framework for one-class classification.** We extend VQ-VAE [1] (architecture above) for semi-supervised learning to leverage a small number of labeled outliers, promoting an efficient learning of ss-VQ-VAE dictionary. While training the network, for the normal/inlier set, $U \in \{X_n\}_{n=1}^{N}$, we apply the $\mathcal{L}_{\text{normal}}$ loss to incorporate the embeddings for normal data in the dictionary. For training with the anomalous/outlier set, $U \in \{Y_m\}_{m=1}^{M}$, we apply the $\mathcal{L}_{\text{anomalous}}$ loss to weed out the embeddings corresponding to the anomalous data. This novel framework is called *ss-VQ-VAE*. In this example, we are using MVTec's bottle (object category) data. During inference, we classify U as normal/anomalous based on the score $\|\mathcal{E}(U;\theta^{\mathcal{E}}) - \mathcal{Q}(U;\theta^{\mathcal{Q}})\|_2$.

Component Analysis) are some refined non-deep-learning-based methods. Deep learning provides a new solution for the feature extraction task which is an important step in anomaly detection. The aim is to learn all the intrinsic features corresponding to the normal class which will later be used for comparison with the test input features for its classification.

While conventionally anomaly detection is solved in the unsupervised regime where the training set contains the data from only the normal/inlier class, there has been some research in the semi-supervised learning regime as well. The autoencoder-based methods aspire to extract the features for the normal class by applying the learning strategies operating in the image space (reconstruction-based losses) and/or the latent space. The common objectives for training unsupervised anomaly detection involve only or a combination of the methods: maximum likelihood, reconstruction error [16], and latent space error [12]. While DeepSVDD [12] is an unsupervised approach, DeepSAD [13] is the corresponding semi-supervised approach for anomaly detection, where the learning occurs in the latent space. The aim is to encapsulate the encoded normal data in a compact hypersphere while ensuring the outlier embeddings lie far away from this hypersphere. While autoencoder-based methods for anomaly detection [16] focus solely on the reconstruction loss, variational autoencoder-based methods [21]–[23] for anomaly detection focus on the reconstruction loss and/or on discriminating the embeddings in the latent space.

The primary motivation for incorporating vector quantization (VQ) in autoencoders is the fact that images contain a lot of redundant information since most of the pixels are correlated. The VQ-VAE [1] architecture discretizes latent space embeddings, and thus can reduce such redundant information. It has currently been used to detect anomalies in the medical domain [5] in an unsupervised learning setup. Our approach incorporates semi-supervision into the VQ-VAE [1] to weed out the anomalous embeddings in the ss-VQ-VAE dictionary. We also design a novel formulation for calculating the anomaly score. Other relevant approaches including FRaC [24] ensembles SVMs, kernel models as well as decision trees, and proposes a surprisal-based anomaly score for the ensembled model. GANomaly [25] presents a deep-learning-based architecture, which learns normal and anomalous features by GAN. Nevertheless, they are hard to train. DifferNet [26] uses a pre-training-based architecture. It is easier to train but heavily relies on pre-trained knowledge. Unlike these methods, ss-gVAE [22] proposes a VAE-based model, which is more efficient for training and does not require pre-training. However, the continuous latent space representation in ss-gVAE [22] still needs complex high dimensional manifold learning, which involves tuning of multiple hyperparameters. Moreover, in the noise and information redundancies in the continuous latent space may lead to incorrect anomaly detection results. In this paper, we propose ss-VQ-VAE to generate discretized latent space to solve this problem. Our ss-VQ-VAE reduces the noise and information redundancies and requires fewer hyperparameters than previous methods. We also present a semi-supervised learning scheme and anomaly score to better train and leverage our discretized latent space for anomaly detection prediction.

III. Methodology

A. Problem Formulation and Method Overview

Given an input image U, we attempt to classify it as normal or anomalous. To this end, we propose ss-VQ-VAE framework to predict an anomaly score $s(U)$ for the input image which tends to be higher for anomalous images and lower for normal images. To train our ss-VQ-VAE, we take a set of inliers (normal images) and a fraction of outliers (anomalous images) for semi-supervision. This is to improve our understanding of discriminating inliers from outliers and learn a good ss-VQ-VAE dictionary. We introduce our ss-VQ-VAE framework with its learning and inference strategies in this section. We provide a symbol table in Table II at the bottom of the paper, containing the notations used throughout the paper along with their explanations for easy reference.

B. ss-VQ-VAE for Anomaly Detection

Let the random field U model an image in the space $\mathcal{U}$. Let a DNN-based encoder model a mapping $\mathcal{E}(.;\theta^{\mathcal{E}})$, parameterized by $\theta^{\mathcal{E}}$, which maps the input image U to the grid embedding $\mathcal{E}(U;\theta^{\mathcal{E}})$ with each embedding of D dimensions (refer Figure 1). We use the vector-quantization-based (VQ-based) dictionary learning algorithm. The latent ss-VQ-VAE dictionary $\mathcal{Q} \in \mathcal{R}^{K\times D}$ consists of K D-dimension embeddings, where $\mathbf{e_k}$ is the k_{th} embedding in the dictionary $\mathcal{Q}$. With the help of ss-VQ-VAE dictionary lookup for the generated grid embedding $\mathcal{E}(U;\theta^{\mathcal{E}})$, we get the corresponding quantized grid embeddings $\mathcal{Q}(U;\theta^{\mathcal{Q}})$. The i_{th} quantized embedding

$$\mathcal{Q}(U;\theta^{\mathcal{Q}})_i = \mathbf{e_k}, \tag{1}$$

where $k = \text{argmin}_j \|\mathcal{E}(U;\theta^{\mathcal{E}})_i - \mathbf{e_j}\|_2$ and $j \in [1, 2, \dots, K]$.

Let a DNN-based decoder model a mapping $\mathcal{D}(.;\theta^{\mathcal{D}})$, parameterized by $\theta^{\mathcal{D}}$, which maps the quantized grid $\mathcal{Q}(U;\theta^{\mathcal{Q}})$ to the reconstruction for the input U, i.e., $\mathcal{D}(\mathcal{Q}(U;\theta^{\mathcal{Q}});\theta^{\mathcal{D}})$. The VQ objective [1] uses the l_2 error to move the ss-VQ-VAE dictionary embedding vectors $\mathbf{e_k} \in \mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_K}$ towards the encoder outputs $\mathcal{E}(U;\theta^{\mathcal{E}})$.

Let $\{X_n\}_{n=1}^N$ represent the normal training set consisting N data points. In the unsupervised scenario, the loss is formulated as

$$\begin{aligned}
&\mathcal{L}_{\text{normal}}(\{X_n\}_{n=1}^N;\theta^{\mathcal{EQD}}) \\
&:= \frac{1}{N}\sum_{n=1}^{N}\Big[\|\mathcal{D}(\mathcal{E}(X_n;\theta^{\mathcal{E}});\theta^{\mathcal{D}}) - X_n\|_2^2 \\
&+ \|\mathbf{sg}[\mathcal{E}(X_n;\theta^{\mathcal{E}})] - \mathcal{Q}(X_n;\theta^{\mathcal{Q}})\|_2^2 \\
&+ \beta\|\mathcal{E}(X_n;\theta^{\mathcal{E}}) - \mathbf{sg}[\mathcal{Q}(X_n;\theta^{\mathcal{Q}})]\|_2^2\Big],
\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{normal}}$ represents the loss formulation for VQ-VAE [1], without any supervision. *sg* denotes *stopgradient operator* [1], which is defined as an identity at forward computation time and used to freeze a part of the network with the previously learned weights. β is the commitment cost, a tuned hyper-parameter to control the ratio of the third term, for encoder learning, in the loss.

In addition to the training set X for the normal class, our ss-VQ-VAE framework leverages a much smaller set of expert-labeled outliers, $\{Y_m\}_{m=1}^M$, consisting M data points. Our motivation is to produce better reconstruction for the normal images than the anomalous ones. The loss for the outlier samples is written as

$$\begin{aligned}
&\mathcal{L}_{\text{anomalous}}(\{Y_m\}_{m=1}^M;\theta^{\mathcal{EQD}}) \\
&:= -\frac{1}{M}\sum_{m=1}^{M}\Big[\|\mathcal{D}(\mathcal{E}(Y_m;\theta^{\mathcal{E}});\theta^{\mathcal{D}}) - Y_m\|_2^2 \\
&+ \|\mathbf{sg}[\mathcal{E}(Y_m;\theta^{\mathcal{E}})] - \mathcal{Q}(Y_m;\theta^{\mathcal{Q}})\|_2^2 \\
&+ \beta\|\mathcal{E}(Y_m;\theta^{\mathcal{E}}) - \mathbf{sg}[\mathcal{Q}(Y_m;\theta^{\mathcal{Q}})]\|_2^2\Big].
\end{aligned} \tag{3}$$

We take the negation of the loss for the anomalous samples in $\mathcal{L}_{\text{anomalous}}$ as we aspire to learn a good ss-VQ-VAE dictionary useful for fine reconstruction of normal samples, weeding out the anomalous embeddings.

In the aforementioned two losses, $\mathcal{L}_{\text{normal}}$ and $\mathcal{L}_{\text{anomalous}}$, in Equations 2 and 3, we have three terms operating on normal and anomalous samples. The first term is the reconstruction loss to optimize the decoder $\mathcal{D}(.;\theta^{\mathcal{D}})$ and the encoder $\mathcal{E}(.;\theta^{\mathcal{E}})$, both. The second term is used to learn the dictionary $\mathcal{Q}(.;\mathcal{Q})$. Here, due to *sg* operator, the gradients will not pass through the encoder $\mathcal{E}(.;\theta^{\mathcal{E}})$, thus disabling the learning for encoder in this term. We add the third term to further optimize the encoder $\mathcal{E}(.;\theta^{\mathcal{E}})$. Similar to the second term, we stop the gradients from passing through the dictionary $\mathcal{Q}(.;\theta^{\mathcal{Q}})$ embeddings during encoder $\mathcal{E}(.;\theta^{\mathcal{E}})$ optimization.

The overall learning objective is now formulated as

$$\begin{aligned}
&\arg\min_{\theta^{\mathcal{EQD}}} \eta\mathcal{L}_{\text{normal}}(\{X_n\}_{n=1}^N;\theta^{\mathcal{EQD}}) \\
&+ (1-\eta)\mathcal{L}_{\text{anomalous}}(\{Y_m\}_{m=1}^M;\theta^{\mathcal{EQD}}),
\end{aligned} \tag{4}$$

where η is the weighing factor for semi-supervision. In this way, we learn our network parameters $\theta^{\mathcal{EQD}}$.

C. Inference Strategy for Anomaly Detection

For inliers, the $\mathcal{L}_{\text{normal}}$ loss promotes the embeddings $\mathcal{E}(X;\theta^{\mathcal{E}})$ to be mapped to the closest possible ss-VQ-VAE dictionary $\mathcal{Q}(.;\theta^{\mathcal{Q}})$ embeddings, making sure the reconstructions are good with the help of individual loss terms in $\mathcal{L}_{\text{normal}}$. However, the outlier learning from $\mathcal{L}_{\text{anomalous}}$ removes embeddings for outlier samples in the dictionary $\mathcal{Q}(.;\theta^{\mathcal{Q}})$. Thus, we can compare the embeddings of the test input U with elements in the dictionary $\mathcal{Q}(.;\theta^{\mathcal{Q}})$. Given a test input U, we classify it as inlier/outlier based on the anomaly score

$$s(U) = \|\mathcal{E}(U;\theta^{\mathcal{E}}) - \mathcal{Q}(U;\theta^{\mathcal{Q}})\|_2. \tag{5}$$

A larger score $s(U)$ means that the embeddings of U are not similar to the embeddings in the dictionary $\mathcal{Q}(.;\theta^{\mathcal{Q}})$, which indicates that U is more likely to be anomalous. We put a threshold ρ (a tuned parameter) on $s(U)$ to classify U as normal/anomalous.

D. Intuition for Anomaly Score Calculation

Ensemble models like FRaC [24], SVMs, kernel models and decision trees proposes a surprisal-based anomaly score. GAN based anomaly detection methods like GANomaly [25] employ a conditional GAN and takes the difference between the latent mappings of the generator and the encoder as the anomaly score. These anomaly scores cannot be used in autoencoder based models like ss-VAE, DeepSAD, ss-DCAE, and our ss-VQ-VAE. Relying on the latent embeddings is more intuitive for such methods. Thus, we propose an anomaly score for our ss-VQ-VAE, which compares the encoder embedding with the discretized embedding in the VQ dictionary which is lower for the normal points and higher for anomalous, as expected for the anomaly score calculation.

IV. Experiments

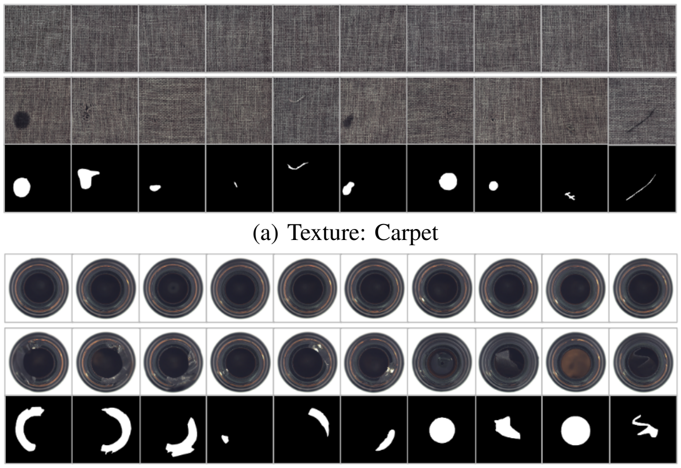
We conduct experiments on real-world datasets: ten texture categories and five object categories of MVTec [27] and the curated industrial inspection corrosion dataset.

A. Competing Methods

We first compare ss-VQ-VAE with its unsupervised version, VQ-VAE, where we take only the normal data points and the $\mathcal{L}_{\text{normal}}$ loss during training. As a deep unsupervised competitor, we take DeepSVDD [12] which aims to enclose the inliers in a compact hypersphere. Under non-deep-learning-based methods, unsupervised OC-SVM [17] is an equivalent of RBF (Radial Basis Function) kernel employed to separate inliers and outliers in the high-dimensional space with a hyperplane. SSAD [28] is a semi-supervised variant of OC-SVM. In order to compare the kernel-based methods with deep learning-based methods, we apply OC-SVM and SSAD to the resulting bottleneck representations from the converged autoencoder in DeepSVDD. We call them as OC-SVM-hybrid and SSAD-hybrid. We also add semi-supervision to DCAE [16] to compare our method with reconstruction-loss-based methods. ss-VAE is a semi-supervised version of variational autoencoders operating on Gaussian distribution. ss-gVAE [22] is a semi-supervised anomaly detection based method which tries to map the data in a generalized-Gaussian domain in order to model the heavy-tailed distributions as well. DifferNet [26] uses a pre-training-based architecture for semi-supervised anomaly detection using adversarial training. It is easier to train than classic GANs but heavily relies on pre-trained knowledge.

B. The MVTec AD Dataset

The MVTec dataset [27] is specifically designed for industrial anomaly detection which includes ten texture categories and five object categories. We take patches of size 64×64 for the texture categories since the anomalies are present in smaller regions. A patch is labeled as anomalous if the corresponding anomaly mask contains at least 5% of anomalous pixels. The dataset contains 2,000 patches for inliers and outliers each in the training set and similarly, 1,000 each in the test set. On the contrary, we take the entire image for classification in objects since the anomaly can only be deciphered in the entire context of the image.



(a) Texture: Carpet

(b) Bottle

Fig. 2. Samples of texture and object categories on the MVTec dataset [27]: Top: images with no anomaly; Middle: images with anomaly; Bottom: anomaly mask corresponding to the images above.

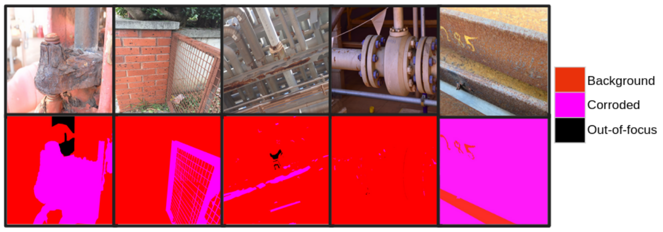


Fig. 3. Corrosion Dataset: Top: images with anomaly (corrosion); Bottom: the mask corresponding to the image above. We consider corroded pixels as anomaly and the rest (Background and Out-of-focus) as normal.

C. The Corrosion Dataset

The studies on automated corrosion detection are hindered due to the lack of public datasets. We have acquired 199 high-resolution images as presented in Figure 3 with the appropriate labels. While fully-supervised methods for corrosion detection (semantic segmentation) require a much larger set of labeled training data to make accurate predictions, our semi-supervised approach is able to identify corrosion without the need of larger datasets. We curate patches of size 512×512 from these high-resolution images. A patch is labeled anomalous if the corresponding anomaly mask contains at least 50% of anomalous pixels. The training set consists of 10,000 patches and the test set has 2,500 patches equally balanced for normal and anomalous classes.

D. Training details

We use the architecture in lines with VQ-VAE [1], consisting encoder $\mathcal{E}$, decoder $\mathcal{D}$, and the ss-VQ-VAE dictionary $\mathcal{Q}$. The architecture details are mapped in Figure 4. The encoder $\mathcal{E}$ consists of two convolutional layers (4×4 filters) followed by another convolutional layer (3×3 filters), and ending with a residual stack consisting a varying number of residual layers. Then, we use a convolutional layer to map the hidden embeddings to the size of the embedding dimension.

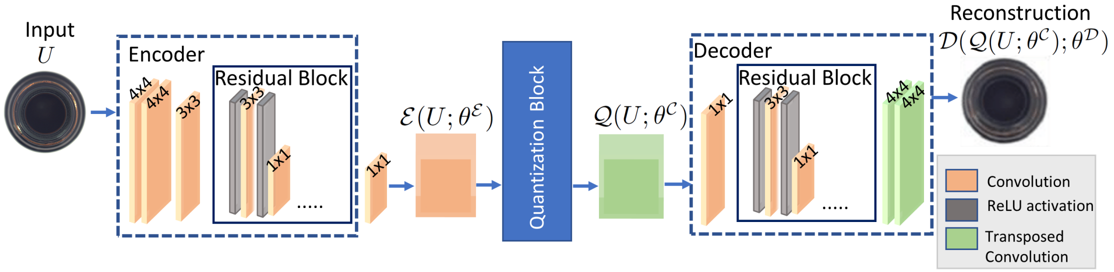


Fig. 4. Architecture details for Semi-Supervised VQ-VAE (ss-VQ-VAE) Learning Framework for OCC.

The decoder network $\mathcal{D}$ consists of a convolutional layer (3×3 filters), and in lines with the encoder $\mathcal{E}$, it contains residual stack followed by two transposed convolutional layers (4×4 filters). We have some hidden layers and residual hidden layers, as per the requirements of dataset. We generally use 128 and 32 hidden and residual hidden layers in our experiments, respectively. The ss-VQ-VAE dictionary $\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$ has K embeddings in it. We set K to 512. D is the dimensionality for each of the latent embeddings in the ss-VQ-VAE dictionary $\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$. In our architecture, we keep $D = 64$. We use the ADAM optimiser [29] with learning rate 1e-3 and evaluate the performance after 15,000 steps with batch-size 128. We use the ADAM optimizer [29] with learning rate 1e-3 and evaluate the performance after 15,000 steps with batch-size 128. We set β and η to 1 and 0.5, respectively. Our method is quite robust to the changes in $\beta \in 0.1, ..., 2.0$ and $\eta \in 10^{-2}, ..., 10^{2}$, as per our sensitivity analysis.

We use the area under the receiver-operating-characteristic curve (AUC) for performance evaluation. For qualitative analysis, we choose the threshold ρ as a point where the difference between $1 - FPR$ and TPR is close to zero in the ROC (Receiver Operating Characteristics) curve, where FPR represents False Positive Rate and TPR represents True Positive Rate. The supervision level $\gamma = M/(M + N)$ is the fraction of outliers used during experiments.

1) Hyperparameter Selection: Hyperparameters tuning plays an important role in ss-VQ-VAE training. $\gamma = M/(M+N)$ is the supervision level for experimental settings. It is the ratio of expert-labeled outlier data provided during the training, where M is the number of anomalous samples and N is the number of normal samples. The value of γ ranges from 0 to 0.5, where $\gamma = 0$ (since $M = 0$) represents an unsupervised scenario while $\gamma = 0.5$ (since $M = N$) denotes a fully-supervised scenario. For sensitivity analysis, we vary values of β ranging from 0.1 to 2.0. We found our method to be quite robust to β like VQ-VAE [1], as the results did not vary for these values of β. Therefore, we generally set $\beta = 1.0$ in our experiments. We have done sensitivity analysis on $\eta \in 10^{-2}, \ldots, 10^{2}$. We find that when $\eta < 0.5$, the performance is highly variable with the change of the supervision level γ. However, when $\eta \geq 0.5$, the performance is quite consistent across all supervision levels. Hence, we generally keep $\eta = 0.5$. In Table 1 and Figure 3, we report AUC results, which do not need the threshold ρ. In t-SNE visualizations (Figure 4), we choose the threshold ρ as a point where the difference between $1 - FPR$ and TPR is close to zero in the ROC (Receiver Operating Characteristics) curve, where FPR represents False Positive Rate and TPR represents True Positive Rate.

E. Results on the MVTec and Corrosion Datasets

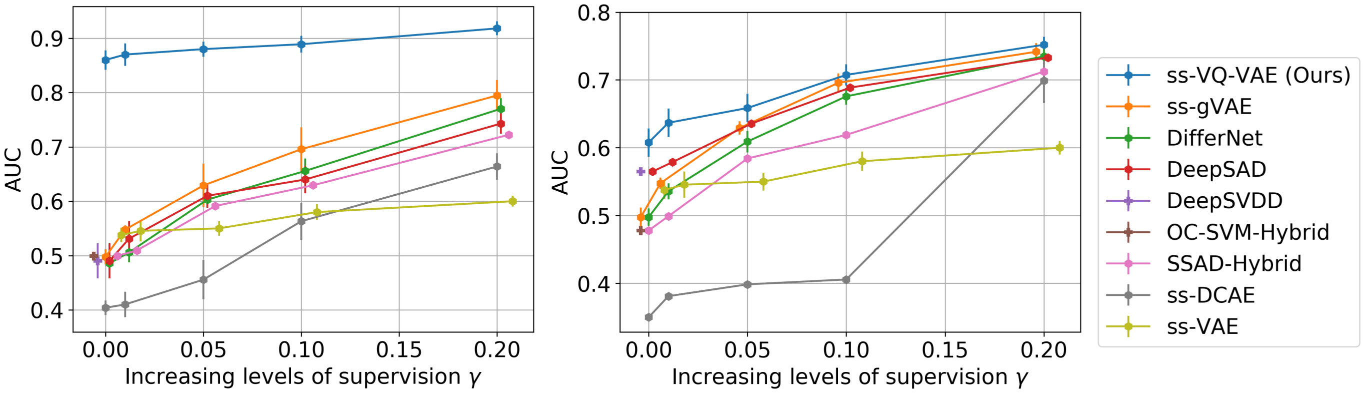
1) Quantitative results: We compare our method with the baselines and ablated versions in Table I for MVTec dataset. DeepSAD, ss-gVAE, ss-VAE, and ss-DCAE are semi-supervised methods using continuous latent embeddings, where ss-gVAE shows the best performance. Compared with ss-gVAE, our method achieves up to 8% improvement on MVTec textures. This demonstrates the significance of using vector quantization based autoencoder approaches to remove redundancies and noise in the latent embeddings. VQ-VAE also uses vector quantization. Our ss-VQ-VAE obtains up to 9% gains over VQ-VAE, which shows the significance of our semi-supervision training framework and our anomaly score. DifferNet is an easier to train version of GANomaly [25] for semi-supervised anomaly detection using adversarial training using pre-training methods but it relies heavily on the pre-training and thus its performance lacks in comparison to our proposed method which is efficient at extracting the features from the training dataset for the normal and differentiates it from the anomalous. Compared to DifferNet, our method achieves up to 3% improvement on MVTec dataset. In Figure 5(a), we plot the comparison for the carpet (texture) category at different supervision levels. As seen in the plot, ss-VQ-VAE performs better than all competing methods at all levels of supervision. On the corrosion dataset, in Figure 5(b), ss-VQ-VAE also shows consistent improvements compared with all baselines at all supervision levels. This proves the effectiveness of our ss-VQ-VAE in the industrial inspection application. Though our method works well with only normal data ($\gamma = 0$), but when providing some supervision, the performance of our method increases resulting in better predictions.

2) Qualitative results: For qualitative comparison we compare our proposed methods' latent representations with the competitive autoencoder based baseline, DeepSAD. In Figure 6, we present t-SNE visualizations on the corrosion dataset and for the carpet category on the MVTec dataset. It shows

TABLE I

AUC results on MVTec (5 Textures, 10 Objects). We compare ss-VQ-VAE with unsupervised VQ-VAE and semi-supervised baselines at $\gamma = 0.2$ supervision. The mean AUC (μ) and standard deviation (σ) show the average and variation in performance across 10 seeds.

Category	**ss-VQ-VAE**	VQ-VAE	ss-gVAE	Deep SAD	SSAD-Hybrid	ss-DCAE	ss-VAE	DifferNet
Carpet	**0.92**	0.86	0.80	0.74	0.72	0.66	0.50	0.77
Grid	**0.76**	0.70	0.72	0.73	0.67	0.59	0.48	0.66
Leather	**0.97**	0.94	0.92	0.94	0.96	0.83	0.54	0.93
Tile	**0.86**	0.65	0.76	0.64	0.74	0.75	0.55	0.95
Wood	0.85	0.74	0.76	0.76	0.67	0.73	0.72	**0.96**
Textures: μ	**0.87**	0.78	0.79	0.76	0.75	0.71	0.56	0.85
Textures: σ	**0.08**	0.12	0.08	0.11	0.12	0.09	0.09	0.13
Bottle	0.87	0.79	0.82	0.77	0.53	0.71	0.53	**0.88**
Cable	0.82	0.69	0.75	0.69	0.58	0.62	0.58	**0.83**
Hazelnut	**0.91**	0.69	0.60	0.57	0.53	0.54	0.55	0.90
Metal nut	**0.76**	0.68	0.75	0.72	0.59	0.45	0.58	0.75
Screw	**0.77**	0.76	0.62	0.61	0.62	0.58	0.58	0.76
Capsule	**0.63**	0.55	0.52	0.51	0.58	0.51	0.51	0.61
Pill	**0.79**	0.72	0.52	0.67	0.59	0.51	0.52	0.65
Toothbrush	0.67	0.59	0.69	**0.70**	0.60	0.54	0.55	0.66
Transistor	**0.84**	0.79	0.67	0.82	0.62	0.46	0.68	0.60
Zipper	**0.63**	0.60	0.52	0.60	0.52	0.55	0.56	0.70
Objects: μ	**0.77**	0.69	0.65	0.67	0.58	0.55	0.56	0.73
Objects: σ	**0.10**	0.08	0.11	0.09	0.04	0.08	0.05	0.11



(a) MVTec's Carpet Dataset **(b) Corrosion Datsaset** **Baselines**

Fig. 5. **Results on the MVTec's Carpet category and Corrosion Dataset.** Comparison of AUC values of our ss-VQ-VAE with baseline methods. The plots (with bars) indicate the variability in AUC across randomly sampled training sets, validation sets, and test sets (10 repeats). The supervision level γ represents the fraction of outliers used in the experiments.

the separability between the inliers and outliers along with the classification accuracy. The classification accuracy is shown as True Positive (Predicted Inlier, Actual Inlier), False Negative (Predicted Outlier, Actual Inlier), True Negative (Predicted Outlier, Actual Outlier), and False Positive (Predicted Inlier, Actual Outlier). Compared with the baseline DeepSAD, our method (ss-VQ-VAE) shows improved separability at all supervision levels. When we increase the level of supervision, the inlier and outlier distributions are getting distinctively separated for both ss-VQ-VAE and DeepSAD. For ss-VQ-VAE, the inlier and outlier clusters are better separated than DeepSAD and the number of false predictions decreases when we increase the level of supervision. There is a clear separation between the inlier and outlier clusters even at 0% supervision for ss-VQ-VAE, which indicates that we obtain good latent embeddings in the dictionary in the unsupervised scenario. By adding supervision, the outlier embeddings are weeded out from the ss-VQ-VAE dictionary, resulting in lesser false predictions.

F. Ablation Study

1) Vector Quantization: In this paper, we propose to use vector quantization for anomaly detection. To verify the effectiveness of vector quantization, we remove it from ss-VQ-VAE (i.e, *ss-VAE*). On the MVTec dataset, the results of ss-VAE are reported in Table I. The mean AUC of 0.56 at 20% supervision signifies that ss-VAE is not able to predict an image as normal or anomalous with confidence. Compared with ss-VAE, our ss-VQ-VAE obtains 31% improvement, which proves the importance of vector quantization.

2) Semi-supervision: From Table I, it can be seen that unsupervised VQ-VAE only shows 0.78 mean AUC for *Texture*, while our ss-VQ-VAE obtains 0.87. Meanwhile, Figures 5 and 6 shows that with the increasing level of supervision, our per-

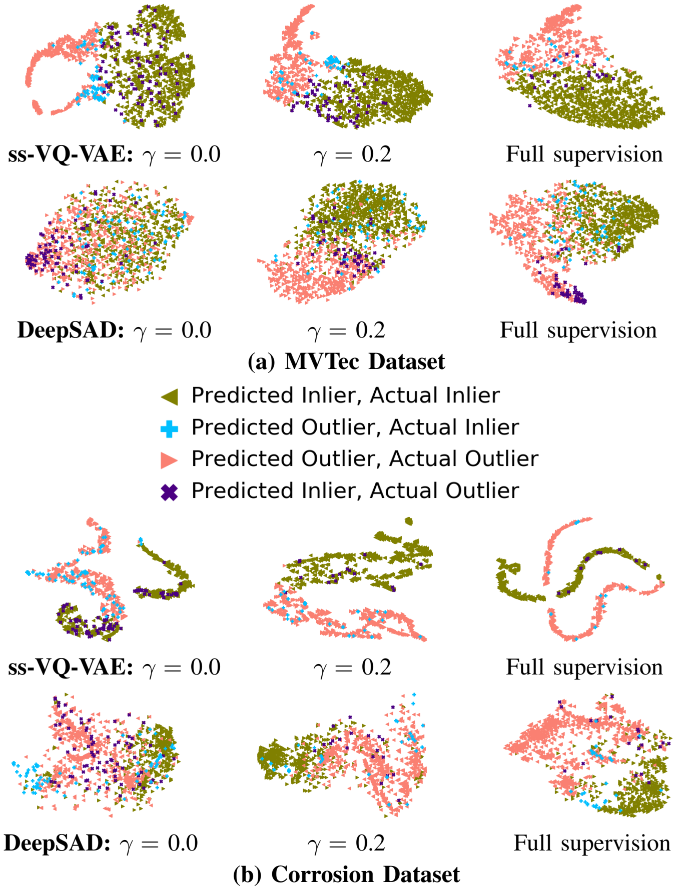


Fig. 6. **Results on the MVTec's carpet category and corrosion dataset:** t-SNE visualizations of latent space distributions on the test set for normal and anomalous images at different supervision level γ.

formance can be further improved. These results demonstrate the effectiveness of semi-supervision in our method.

3) Loss function: We verify the effectiveness of each term in our loss functions. (i) **A1:** We remove the first term (reconstruction term) from our losses. It results in a learning objective equivalent to our baseline DeepSAD. Our ss-VQ-VAE achieves significant improvement over DeepSAD, as seen in Table I and Figure 5, which shows the significance of the reconstruction term. (ii) **A2:** We remove the second and third terms from our losses. These two terms are interdependent, hence we are removing both the terms at the same time to get an intuitive ablated version. The resulting learning objective is equivalent to ss-DCAE. The results in Table I and Figure 5, comparing ss-DCAE with ss-VQ-VAE, prove the importance of the dictionary and encoder learning terms in our losses.

V. Conclusion

In this work, we have introduced ss-VQ-VAE, a novel vector quantized variational autoencoder, incorporating semi-supervision and an anomaly score for anomaly detection with discrete latent space embeddings. We apply semi-supervision with the help of our novel loss formulation in the ss-VQ-VAE dictionary learning. We also introduce the associated anomaly score formulation used to infer an image as anomalous/normal. The proposed approach has merit in assessing anomalies that are present as small features for images at large scales. This is because ss-VQ-VAE is trained by looking at a large fraction of normal and a few anomalous images (as per the level of supervision), and small variations from training data can be detected, in a semi-supervised manner. Conversely, in the real-world scenario, for supervised learning, there is no dataset that is labeled at the large scale (like the case of corrosion detection), and human/expert training is restricted in terms of the level of detail that can be labeled. Extensive experiments on the MVTec and corrosion datasets with comparisons to eight baselines and the ablation study prove the benefits of our proposed ss-VQ-VAE.

Acknowledgement

We wish to extend our gratitude to Woodside-Monash FutureLab at Monash University for generously providing us with the corrosion dataset and appreciate the consistent guidance from Lee Djumas and Maria Garcia De La Banda. We also acknowledge the valuable contributions made by William Nash, Matthias Klapperstueck and Andrew A. Lockwood to our project on corrosion detection.

References

[1] A Van Den Oord, O Vinyals, et al., "Neural Discrete Representation Learning," *Adv. in Neu. Inf. Proc. Sys.*, vol. 30, 2017.

[2] C Hu, K Chen, and H Shao, "A semantic-enhanced method based on deep svdd for pixel-wise anomaly detection," in *Int. Conf. on Mul. and Exp.*, 2021, pp. 1–6.

[3] S Chen, Y Liu, C Liu, TP Chen, and YF Wang, "Domain-generalized textured surface anomaly detection," in *Int. Conf. on Mul. and Exp.*, 2022, pp. 01–06.

[4] L Nie, L Zhao, and K Li, "Glad: Global and local anomaly detection," in *Int. Conf. on Mul. and Exp.*, 2020, pp. 1–6.

[5] SN Marimont and G Tarroni, "Anomaly Detection Through Latent Space Restoration using Vector Quantized Variational Autoencoders," in *Int. Symp. on Bio. Ima.*, 2021, pp. 1764–1767.

[6] M Kimura and T Yanagihara, "Anomaly Detection using GANs for Visual Inspection in Noisy Training Data," in *Asi. Conf. on Comp. Vis*, 2018, pp. 373–385.

[7] C Ma, Z Miao, M Li, S Song, and MH Yang, "Detecting Anomalous Trajectories via Recurrent Neural Networks," in *Asi. Conf. on Comp. Vis*, 2018, pp. 370–382.

[8] P Mantini, Z Li, et al., "A Day on Campus–An Anomaly Detection Dataset for Events in a Single Camera," in *Asi. Conf. on Comp. Vis*, 2020.

[9] Hugh L. Kennedy, "Whitening pre-filters with circular symmetry for anomaly detection in hyperspectral imagery," in *Dig. Ima. Comp.: Tech. and App. DICTA*. 2018, pp. 1–8, IEEE.

[10] SD Bhattacharjee, J Yuan, Z Jiaqi, and Y Tan, "Context-aware graph-based analysis for detecting anomalous activities," in *Int. Conf. on Mul. and Exp.*, 2017, pp. 1021–1026.

[11] Li-Li Wang, H. Y. T. Ngan, W. Liu, and N. H. C. Yung, "Anomaly detection for quaternion-valued traffic signals," in *Dig. Ima. Comp.: Tech. and App. DICTA*. 2016, pp. 1–4, IEEE.

[12] L Ruff et al., "Deep One-Class Classification," in *Int. Conf. on Mac. Lear.*, 2018, pp. 4393–402.

[13] L Ruff et al., "Deep Semi-Supervised Anomaly Detection," in *Int. Conf. on Lea. Rep.*, 2020.

[14] Y Xia, X Cao, F Wen, G Hua, and J Sun, "Learning Discriminative Reconstructions for Unsupervised Outlier Removal," in *Int. Conf. on Comp. Vis.*, 2015, pp. 1511–9.

[15] R Chalapathy, AK Menon, and S Chawla, "Robust, Deep and Inductive Anomaly Detection," in *Mac. Lear. and Know. Dis. in Dat., ECML PKDD*, 2017, pp. 36–51.

TABLE II
Symbol table: EXPLANATIONS FOR THE NOTATIONS USED

Symbol	**Explanation**
X	Input image/random field for normal samples, every image is represented as a random field
Y	Input image/random field for abnormal samples
$\mathcal{E}(\cdot;\theta^{\mathcal{E}})$	Encoder for VQ-VAE, maps the input (X or Y) to the continuous latent embedding, $\mathcal{E}(X;\theta^{\mathcal{E}})$
$\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$	ss-VQ-VAE dictionary for VQ-VAE, maps the continuous latent embedding, $\mathcal{E}(X;\theta^{\mathcal{E}})$, to the quantized embedding $\mathcal{Q}(X;\theta^{\mathcal{Q}})$
$\mathcal{D}(\cdot;\theta^{\mathcal{D}})$	Decoder for VQ-VAE, maps the discrete latent embedding, $\mathcal{Q}(X;\theta^{\mathcal{Q}})$, to the reconstruction $\mathcal{D}(\mathcal{Q}(X;\theta^{\mathcal{Q}});\theta^{\mathcal{D}})$ which is very close to input for normal sample but far away for anomalous sample
$\theta^{\mathcal{E}}$	Parameters of the encoder $\mathcal{E}(\cdot;\theta^{\mathcal{E}})$ for ss-VQ-VAE, maps the input (as normal X or abnormal Y) to the latent space (as latent vector $\mathcal{E}(X;\theta^{\mathcal{E}})$)
$\theta^{\mathcal{D}}$	Parameters of the decoder $\mathcal{D}(\cdot;\theta^{\mathcal{D}})$ for ss-VQ-VAE, maps the embedded input (in the latent space) to the reconstructed space
$\theta^{\mathcal{Q}}$	Parameters of the ss-VQ-VAE dictionary $\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$ for ss-VQ-VAE
$\theta^{\mathcal{E}\mathcal{Q}\mathcal{D}} := \theta^{\mathcal{E}} \cup \theta^{\mathcal{Q}} \cup \theta^{\mathcal{D}}$	ss-VQ-VAE parameters; combination of the parameters for encoder $\mathcal{E}(\cdot;\theta^{\mathcal{E}})$, ss-VQ-VAE dictionary $\mathcal{Q}(\cdot;\theta^{\mathcal{Q}})$, and decoder $\mathcal{D}(\cdot;\theta^{\mathcal{D}})$
$\mathcal{L}_{\text{normal}}(X,\theta^{\mathcal{E}\mathcal{Q}\mathcal{D}})$	Loss for normal samples in ss-VQ-VAE or the unsupervised learning with VQ-VAE (since there is no semi-supervision introduced here)
$\mathcal{L}_{\text{anomalous}}(Y,\theta^{\mathcal{E}\mathcal{Q}\mathcal{D}})$	Loss for anomalous samples in ss-VQ-VAE
M	Total number of normal samples
N	Total number of anomalous samples
γ	Supervision level, $\gamma = M/(M+N)$. It varies from 0 to 0.2 in our experiments. $\gamma = 0.5$ represents full-supervision.
sg	Learnable stopgradient operator, which is defined as an identity at forward computation time and used to freeze a part of the network with the previously learned weights.
β	Coefficient for the third term in the $\mathcal{L}_{\text{normal}}$ and $\mathcal{L}_{\text{anomalous}}$ losses used to learn the encoder part of the architecture when keeping the weights for ss-VQ-VAE dictionary constant. We generally keep $\beta = 1$.
η	Supervision coefficient in the learning objective. It weighs the losses corresponding to normal and anomalous samples. We generally keep $\eta = 0.5$ to give equal weights to both loss terms.
ρ	Threshold for our anomaly score. For qualitative analysis, we choose it as a point at which the difference between $1 - FPR$ and TPR is close to zero in the ROC curve, where FPR represents False Positive Rate and TPR represents True Positive Rate.

[16] J Masci, U Meier, D Cireşan, and J Schmidhuber, "Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction," in *Int. Conf. on Art. Neu. Net.*, 2011, pp. 52–9.

[17] B Scholkopf, J Platt, J Shawe-Taylor, A Smola, and R Williamson, "Estimating the Support of a High-dimensional Distribution," *Neural Comp.*, vol. 13, no. 7, pp. 1443–71, 2001.

[18] D Tax and R Duin, "Support Vector Data Description," *Mac. Lear.*, vol. 54, no. 1, pp. 45–66, 2004.

[19] B Scholkopf, A Smola, and K Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[20] H Hoffmann, "Kernel PCA for Novelty Detection," *Pat. Reco.*, vol. 40, no. 3, pp. 863–74, 2007.

[21] D Kingma and M Welling, "Auto-Encoding Variational Bayes," in *Int. Conf. on Lear. Rep.*, 2014.

[22] R Sharma, S Mashkaria, and SP Awate, "A Semi-Supervised Generalized VAE Framework for Abnormality Detection using One-Class Classification," in *Wint. Conf. on App. of Comp. Vis.*, 2022, pp. 595–603.

[23] R Sharma and SP Awate, "Robust and Uncertainty-Aware VAE (RU-VAE) for One-Class Classification," in *Int. Symp. on Bio. Imag.*, 2022, pp. 1–5.

[24] K Noto, C Brodley, and D Slonim, "Frac: a feature-modeling approach for semi-supervised and unsupervised anomaly detection," *Data Min. and Know. Dis.*, vol. 25, pp. 109–133, 2012.

[25] S Akcay, A Atapour-Abarghouei, and T P Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asi. Conf. on Comp. Vis.*, 2019, pp. 622–637.

[26] M Rudolph, B Wandt, and B Rosenhahn, "Same Same But Differnet: Semi-supervised Defect Detection with Normalizing Flows," in *Win. Conf. on App. of Comp. Vis.*, 2021, pp. 1907–1916.

[27] P Bergmann, M Fauser, D Sattlegger, and C Steger, "MVTec AD–A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," in *Conf. on Comp. Vis. and Pat. Reco.*, 2019, pp. 9592–600.

[28] N Görnitz, M Kloft, K Rieck, and U Brefeld, "Toward Supervised Anomaly Detection," *Jou. of Art. Int. Res.*, vol. 46, pp. 235–262, 2013.

[29] D Kingma and B Jimmy, "Adam: A Method for Stochastic Optimization," *Int. Conf. on Lear. Rep.*, 2015.