Leveraging Vector-Quantized Variational Autoencoder Inner Metrics for Anomaly Detection

Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, Sébastien Lefèvre

Leveraging Vector-Quantized Variational Autoencoder Inner Metrics for Anomaly Detection

Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, Sébastien Lefèvre
IRISA, Université Bretagne Sud, UMR 6074
56000 Vannes, France
{hugo.gangloff, minh-tan.pham, luc.courtrai, sebastien.lefevre}@irisa.fr

***Abstract*—Anomaly Detection (AD) is an important research topic, with very diverse applications such as industrial defect detection, medical diagnosis, fraud detection, intrusion detection, etc. Within the last few years, deep learning-based methods have become the standard approach for AD. In many practical cases, the anomalies are unknown in advance. Therefore, most of challenging AD problems need to be addressed in an unsupervised or weakly supervised framework. In this context, deep generative models are widely used, in particular Variational Autoencoder (VAE) models. VAEs have been extended to Vector-Quantized VAEs (VQ-VAEs), a model increasingly popular because of its versatility enabled by the discrete latent space. We present for the first time a robust approach which takes advantage of the inner metrics of VQ-VAEs for AD. We show that the distance between the output of the encoder and the codebook vectors of a VQ-VAE provides a valuable information which can be used to localize the anomalies. In our approach, this metric complements a reconstruction-based metric to improve AD results. We compare our model with state-of-the-art AD models on three standards datasets, including the MVTec, UCSD-Ped1 and CIFAR-10 datasets. Experiments show that the proposed method yields high competitive results.**

I. Introduction

A. Anomaly Detection

Anomaly Detection (AD) is a field of research which has been of interest for several decades, with a wide variety in applications [1]. As defined by [2], an anomaly is an observation that highly deviates from other observations, as much as to arouse suspicion that it was generated by a different mechanism. In other words, the anomalous observation deviates from some underlying concept of normality. As many other fields, it has been revolutionized by deep learning approaches which have yielded new state-of-the-art results thanks to the unprecedented possibilities of capturing and modeling the normality [3].

AD in an unsupervised context is the most common approach in practical cases. Indeed, anomalies are unknown in advance, hence the impossibility to gather labeled anomalous data to train the deep model. From now on, our work will focus on unsupervised AD with images. Following [4], there are three main approaches (often mixed in practice) for unsupervised AD. The detection task can be:

- feature-extraction-based, which relies on a distance in the feature space [5], [6];
- probability-based, which makes use of distributions or statistical tests to detect anomalies [7], [8];
- reconstruction-based, which computes distances between inputs and reconstructions [9], [10].

Reconstruction-based methods are the most explored ones in the literature. They are often based on deep generative latent variable models called Variational Autoencoders (VAEs) which we now present. Note that our new approach presented in this paper can be seen as a combination of a reconstruction-based and a feature-extraction-based approaches.

Remark: When working with AD on images, it is significant to distinguish two tasks: image-wise AD and pixel-wise AD. Image-wise AD involves the detection of the anomalous image as a whole. On the other hand, pixel-wise AD corresponds to the localization of each anomaly. Both cases will be addressed in this article.

B. Related work: anomaly detection with VAE-like models

In the unsupervised and weakly supervised AD contexts, deep generative models are a popular choice of models [11]. Among these models, particular deep latent variable models called Variational Autoencoders (VAEs) [12] have been widely used. They are defined in a probabilistic framework detailed in [13]. In a nutshell, VAEs transform an input image $\boldsymbol{x}$ into a compressed representation $\boldsymbol{z}$ through a stochastic encoder network $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$. The image is then reconstructed to form the reconstruction $\hat{\boldsymbol{x}}$, by sampling from a stochastic decoder network $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. The model is trained, for both network parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$, by minimizing the Evidential Lower Bound which reads:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x}) = \underbrace{\mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]}_{\text{reconstruction term}} - \underbrace{\mathbb{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}))}_{\text{regularization term}}. \quad (1)$$

Intuitively, the first term encourages the reconstruction $\hat{\boldsymbol{x}}$ to be similar to the input $\boldsymbol{x}$ under the constraint of a regularization term under the form of Kullback-Leibler (KL) divergence. It can be seen that VAEs give access to several metrics computed from the continuous latent space and the reconstructions. Such metrics form the foundation of many image-wise and pixel-wise AD approaches. Some of these metrics are, for example: the residual images between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, the evaluation the reconstruction term and/or of the KL term of Equation 1, the derivative of Equation 1 with respect to $\boldsymbol{x}$, etc. Many combinations of these ideas have been successfully applied

to weakly supervised AD giving rise to an important number of recent studies such as [9], [10], [14], [15], [16].

Recently, Vector-Quantized VAEs (VQ-VAEs) [17], [18] have introduced the idea of a discrete latent space. We present them in details in Section II-A. In the literature, several recent studies have already reported use of VQ-VAEs for AD (*e.g.* [4], [8], [19]). However, these works focus on much more complex models and, in particular, they introduce, in a second step, autoregressive models trained on the latent space. Comparisons with such approaches are therefore out of scope of our paper. Indeed, our main contribution is to show that VQ-VAEs are able to learn a latent representation of the normality which provides robust and competitive metrics for AD, without additional complexity, in a similar way of the best results obtained with VAEs. Indeed, while the VAE metrics have been extensively studied, we aim at harnessing the inner metrics of the VQ-VAEs for AD, and then at comparing VQ-VAEs with VAEs on this precise point. To the best of our knowledge, this is the first effort to conduct such a study in the field of AD.

C. Organization of the article

Our paper is organized as follows. In the next section, we introduce VQ-VAEs and how they can be used to address AD problems. We then present a new and robust approach which is based on the inner metrics of the VQ-VAEs and which aims at AD both at the image and pixel levels, via the computation of an anomaly map. In the last section, we compare our approach with some other state-of-the-art approaches on the MVTec [20], UCSD-Ped1 [21] and CIFAR-10 [22] datasets for, respectively, the detection of anomalies on industrial images, the detection of anomalies on crowd monitoring images, and the classification of anomalous real-world images.

II. Vector-Quantized Variational Autoencoders

A. The model

VQ-VAEs, first introduced in [17], are models which include discrete latent variables. They can learn rich, yet compressed, latent representations and have been used to produce much sharper reconstruction than traditional VAEs [18], [23]. Therefore, this makes VQ-VAEs an interesting model for AD, suggesting much less noisy residual images in reconstruction-based approaches, for example.

To cope with the discrete latent space, VQ-VAEs are trained differently from standard VAEs. In particular, the encoder becomes deterministic while the decoder remains stochastic. We now denote the encoder by $\mathrm{Enc}_{\boldsymbol{\varphi}}$. Let M be the number of possible states for the latent variable $z_k, \forall k \in \{1, \dots, K\}$, K being the dimension of the latent space. VQ-VAEs integrate a codebook, *i.e.*, a set of vectors $(e_1, \dots, e_M)$, each one in $\mathbb{R}^D$, with D a positive integer. From the encoder output, $\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}$, we choose the closest codebook vector for $z_k, \forall k$, following a deterministic decision, *i.e.*,

$$z_k = \underset{m \in \{1,\dots,M\}}{\arg\min} \|(z_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})})_k - e_m\|_2. \tag{2}$$

This can be associated to a deterministic and categorical encoding distribution as follows:

$$q_{\boldsymbol{\varphi}}(z_k = m|\boldsymbol{x}) = \begin{cases} 1 \text{ if } m = \underset{m \in \{1,\dots,M\}}{\arg\min} \|(z_{\mathrm{Enc}_{\boldsymbol{\varphi}}})_k - e_m\|_2, \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

The inputs of the decoder, denoted by $\boldsymbol{z}_{\mathrm{Dec}_{\boldsymbol{\theta}}}$, are then also deterministically set as $(z_{\mathrm{Dec}_{\boldsymbol{\theta}}})_k = e_{z_k}, \forall k$. Note that, as done in this article, when processing images with convolutional encoders and decoders, the latent space $\boldsymbol{z}$ can also be convolutional, *i.e.*, it consists of a latent image.

The loss function is then composed of the same reconstruction term as in standard VAEs but also of a squared $\ell 2$ term to ensure that the codebook is learnt. Indeed, because of the deterministic operations, the gradient can not flow from the decoder input to the encoder output: it has to be automatically copied; thus bypassing the codebook which cannot be updated. This second term is called the codebook *alignment term*. A third regularizing term is also added in the loss for stability of the training procedure, which leads to the VQ-VAE loss, for an image $\boldsymbol{x}$:

$$\mathcal{L}^{VQ-VAE}_{\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{e}}(\boldsymbol{x}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{Dec}_{\boldsymbol{\theta}}(\boldsymbol{x})}) + \underbrace{\|\mathrm{sg}[\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}] - \boldsymbol{e}\|_2^2}_{\text{alignment term}} + \beta \|\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})} - \mathrm{sg}[\boldsymbol{e}]\|_2^2, \tag{4}$$

with β a scalar parameter and sg the stop gradient operator.

B. Anomaly Detection with VQ-VAEs

Similar to the metrics used for VAEs, we now define metrics for AD with VQ-VAEs. Since we are interested in pixel-wise AD, following [24], we define a reconstruction-based anomaly map, called SM, which uses the Structural Similarity Index Measure (SSIM) [25] measure. For each pixel i, we have:

$$\mathrm{SM}(x_i) = \mathrm{SSIM}(\boldsymbol{p}_i, \boldsymbol{q}_i) = \frac{(2\mu_{\boldsymbol{p}}\mu_{\boldsymbol{q}} + c_1)(2\sigma_{\boldsymbol{pq}} + c_2)}{(\mu_{\boldsymbol{p}}^2 + \mu_{\boldsymbol{q}}^2 + c_1)(\sigma_{\boldsymbol{p}}^2 + \sigma_{\boldsymbol{q}}^2 + c_2)}, \tag{5}$$

where $\boldsymbol{p}_i$ (resp. $\boldsymbol{q}_i$) is a patch around pixel i of $\boldsymbol{x}$ (resp. $\hat{\boldsymbol{x}}$). μ_p, σ_p and σ_{pq} represent, respectively, the mean, the standard deviation and the covariance of the patches. The scalars are set to $c_1 = 0.01$ and $c_2 = 0.03$ [25].

We also define a new metric producing a latent space anomaly map, intrinsic to VQ-VAEs, which we call the Alignment Map (AM):

$$\mathrm{AM}(\boldsymbol{x}) = \|\mathrm{sg}[\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}] - \boldsymbol{e}\|_2^2. \tag{6}$$

The intuition behind the AM anomaly map is as follows. During training time, the codebook vectors are trained to be close to the output of the encoder, and reciprocally, in virtue to the last two terms in Equation 4. Therefore, at testing time, the anomalies (encoded in $\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}$), which have not been seen yet by the model, will be far from the codebook vectors, relatively to the normal features (also encoded in $\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}$). This is what is reflected in the AM.

It should be noted that this particular anomaly map is defined in the latent space and is not directly usable. In the

next section, we construct an approach to efficiently use the AM to segment anomalies.

Remark: The AM resonates with the approaches which use the Kullback-Leibler divergence value of a VAE as a metric to localize anomalies [10]. However, as our experiments will show in Section III, the alignment loss of a VQ-VAE seems to be a much more robust and more interpretable way to localize anomalies as our results are competitive to those yielded by the state-of-the-art approaches.

C. Improved Anomaly Detection using the Alignment Map

The AM is a small image with same dimension as the latent space where some pixels stand out. Those pixels can be seen as *markers*. They correspond to the latent variables with high alignment loss, *i.e.*, anomalies. To be used in addition to the SM anomaly map, the AM is first upsampled. Then it undergoes a morphological grey dilation which aims at emphasizing the markers. However, none of these markers correctly represent the anomaly since they are just upsampled pixels. Therefore, we propose to recover a more realistic shape for the anomaly by multiplying the AM with the SM. Figure 1 graphically summarizes all the steps of our approach which we call VQ-VAE SSIM+AM.

One of the advantages of this approach is that it does not rely solely on the reconstruction of the model. Indeed, traditional VAE approaches rely on the promise that anomalies, unseen as training will disappear at the output; and they are then isolated based on the principle of the residual image [9]. However, because of the intrinsic blurriness of the reconstructions, this often stands out as a very complex task. Therefore, the AM, as a source of information to localize the anomalies, is useful insofar as it does not directly rely on the reconstructions.

III. Experiments & Results

A. Network architecture

The same VQ-VAE architecture is used in the following experiments. It is based on the original VQ-VAE architecture [17]:

- The encoder consists in three convolutional layers (kernel size 4, stride 2 and padding 1), each followed by a ReLU activation and Batch Normalization. It is then followed by three residual layers (decomposed as a ReLU activation, a convolutional layer (kernel size 3, stride 1, padding 1), a ReLU activation and a convolutional layer (kernel size 1, stride 1, padding 0)). All the depth dimensions are 256, except the input images which have depth 1 or 3.
- The latent space image has the width and height of the original image divided by 8 (when 3 convolutions described above are stacked in the encoder). The codebook size is set to $M = 512$ and each vector of the codebook has dimension $D = 256$.
- The decoder is constructed as the reverse of the encoder.

Moreover, when the image pixels are encoded in $[0, 255]$. We rescale the input images into the range $[0, 1]$ and use a decoder where we consider $\boldsymbol{x}$ as the realization of a continuous Bernoulli random variable [26]. We found that this provided much better results and reduced convergence issues while training the VQ-VAE, as compared to considering a Gaussian distribution for the stochastic decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$.

The VQ-VAE model is tested against comparable approaches, *i.e.*, approaches which are designed in a similar way. The approaches we selected from the literature are then solely based on an encoder-decoder architecture and inner metrics, without any additional modules and very limited pre- and post-processing.

Remark: In the following experiments, training the models does not make use, in any way, of labeled data. However, because we train the models only on normal data, devoid of anomalies, this is, strictly speaking, not an *unsupervised* context but rather a *weakly supervised* context.

B. MVTec dataset

The MVTec dataset [20] is a standard dataset for AD on images. The dataset provides non-anomalous and defective RGB images of manufactured objects from 15 different categories. Several defect types are available for each category. In this section, we resize the image to 256×256 pixels, leading to a 32×32 latent space. Following the literature, we give the results in terms of pixel-wise ROCAUC which is generated from the heatmap available after processing and the available ground truth. The full approach for AD described in Section II-C is called VQ-VAE SSIM+AM. We also have the VQ-VAE SSIM approach where the SM anomaly map is directly used to compute the ROCAUC score, without using the AM. We compare with the ROCAUC scores of a classical Autoencoder with SSIM (AE SSIM) [20], the state-of-the-art Visually Explained VAE (VEVAE) [14] and the state-of-the-art Fully Convolutional Data Description (FCDD) [27].

Table I shows the score for the models over all the categories of the dataset. The scores for the other methods are taken from

Category	AE SSIM [20]	VEVAE [14]	FCDD [27]	VQ-VAE SSIM	VQ-VAE SSIM+AM
Carpet	0.87	0.78	**0.96**	0.92	0.94
Grid	0.94	0.73	0.91	**0.99**	**0.99**
Leather	0.78	0.95	**0.98**	**0.98**	**0.98**
Tile	0.59	0.80	**0.91**	0.70	0.75
Wood	0.73	0.77	**0.88**	0.82	0.84
Bottle	0.93	0.87	**0.97**	0.94	0.95
Cable	0.82	**0.90**	**0.90**	0.87	0.87
Capsule	**0.94**	0.74	0.93	0.93	**0.94**
Hazelnut	0.97	0.98	0.95	0.98	**0.99**
Metal Nut	0.89	0.90	**0.94**	0.89	0.90
Pill	**0.91**	0.83	0.81	0.86	0.90
Screw	0.96	0.97	0.86	**0.98**	**0.98**
Toothbrush	0.92	0.94	0.94	0.96	**0.97**
Transistor	0.90	**0.93**	0.88	0.77	0.78
Zipper	0.88	0.78	0.92	0.97	**0.98**
Mean	0.86	0.86	**0.92**	0.90	**0.92**

TABLE I: ROCAUC scores on the MVTec dataset.

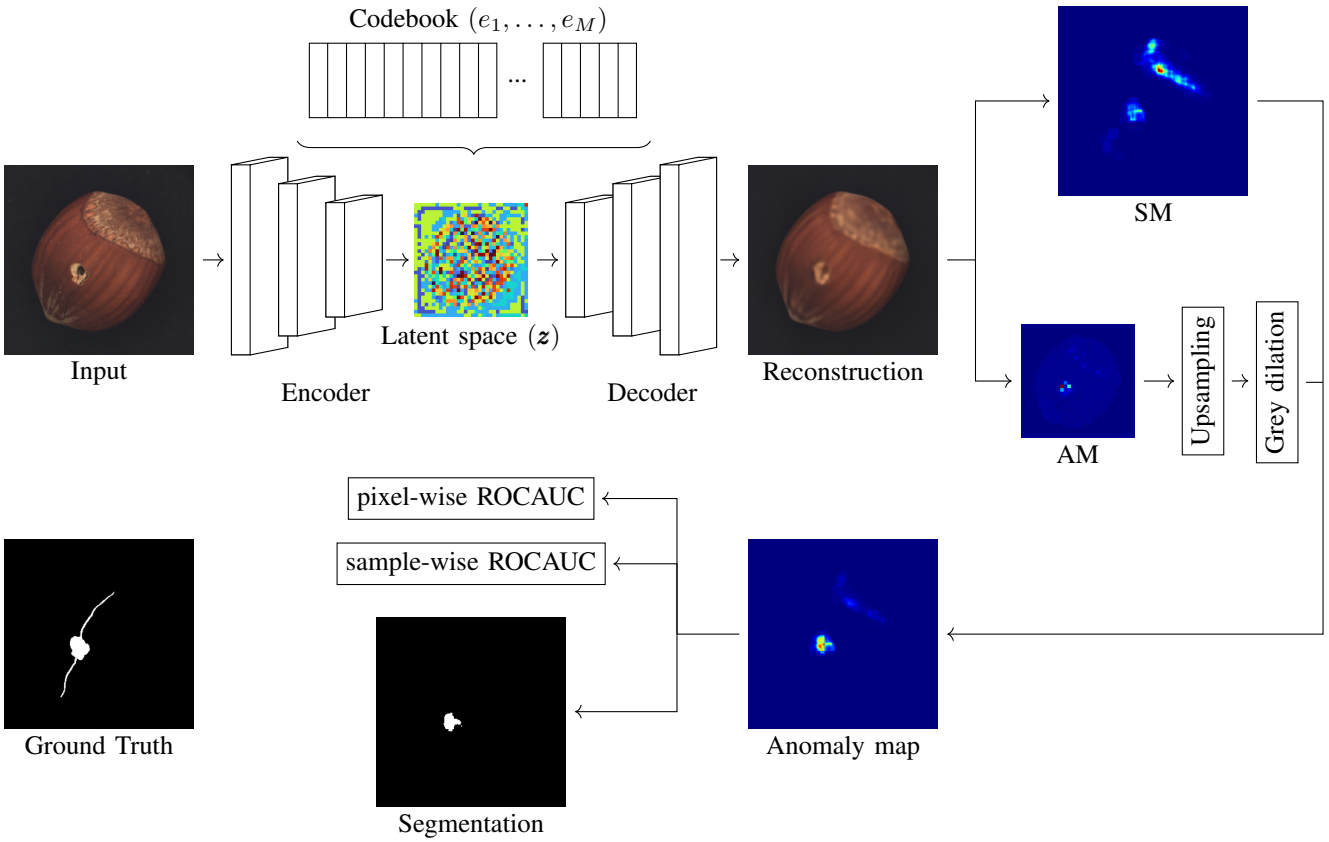


Fig. 1: Illustration of the proposed workflow for improved AD with VQ-VAE (Sections II-B and II-C). The network architecture is described in Section III-A. This approach is called VQ-VAE SSIM+AM. It is possible to directly use the SM anomaly map for ROCAUC computations or anomaly segmentation. In this case the approach is called VQ-VAE SSIM.

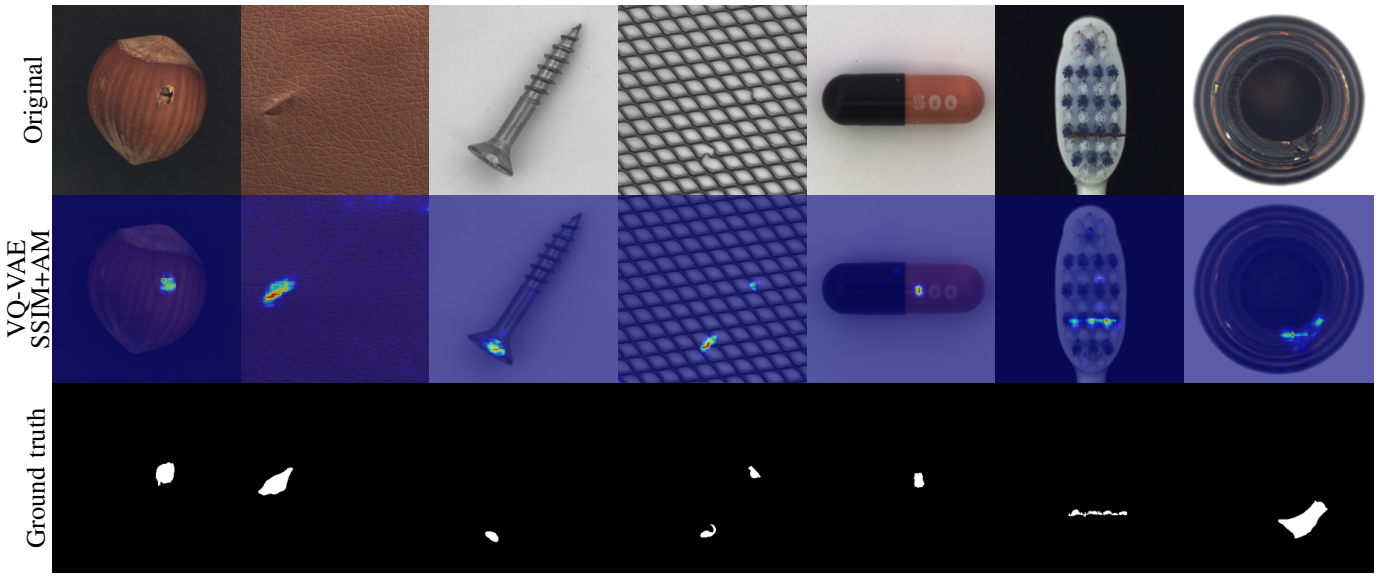


Fig. 2: Selected illustrations for the MVTec experiment. Anomaly maps and reconstructions from our proposed approach. *Top row:* original image. *Middle row:* the anomaly maps overlaying the VQ-VAE reconstruction. *Bottom row:* The segmented anomaly.

	VAE L2 [9]	VEVAE [14]	VQ-VAE SSIM	VQ-VAE SSIM+AM
ROCAUC	0.86	0.92	**0.95**	**0.95**

TABLE II: ROCAUC scores on the UCSD-Ped1 dataset.

their respective referenced paper. We can see that our VQ-VAE SSIM+AM approach performs similarly as the FCDD approach which represents the state-of-the-art result on MVTec for this family of approaches. Our proposed approach also gives better results than AE SSIM and VEVAE. This suggests that VQ-VAEs might be more robust baseline architectures than AEs and VAEs. VQ-VAE SSIM also performed worse than VQ-VAE SSIM+AM, which highlights the interest of using the additional information provided by the AM to improve the final anomaly map.

We noticed that the worst results of the VQ-VAE approaches seem to be linked with relatively big defects being too well reconstructed, for example, a black spot over a background composed of black spots (in *Tile*) or a missing part of an object which reveals more background (in *Transistor*). In these cases, the metrics failed to localize the anomalies. On the other hand, the strength of the model seems to reside in its ability to detect even the smallest defects thanks to the sharpness of the reconstructions permitted by VQ-VAEs (holes in *Hazelnut* or small damages in *Screw*). Figure 2 provides some selected graphical illustrations of the experiment.

C. UCSD-Ped1 dataset

This second experiment addresses another standard dataset for AD on images and videos: the UCSD-Ped1 dataset [21]. The dataset is composed of black and white video sequences of pedestrians walking in a park. This represents $6,400$ images for training and $2,000$ images for testing. In this experiment, we resize the image to 128×128 as done in [14]. Following the literature, the metric used will be the pixel-wise ROCAUC. In the context of this experiment, localizing the anomalies consists in localizing all non-pedestrian moving objects in this park scenery (car, bikes, skateboards, etc.).

We compare our VQ-VAE-based approaches with a classical VAE architecture [9] and the VEVAE [14] architecture whose results are available in the literature. Note that in our approach, because of the small resolution of the original images, the VQ-VAE encoder and decoder described in Section III-A are reduced to two convolutional layers. The latent space then has dimension 32×32. Moreover, we found out that much better results were obtained by performing a morphological grey dilation also on the SM anomaly map. This is reflected in the illustrations of Figure 3, where patches from the SSIM computation are emphasized.

Table II gives the ROCAUC for the models over all the testing dataset. The scores for the other models are both taken from [14]. We can see an improvement over the other models, which signifies that more accurate anomaly maps are produced by the VQ-VAE approaches. In the context of this experiment, it is not possible to favor one of our VQ-VAE versions. Illustrations of the experiment are available in Figure 3.

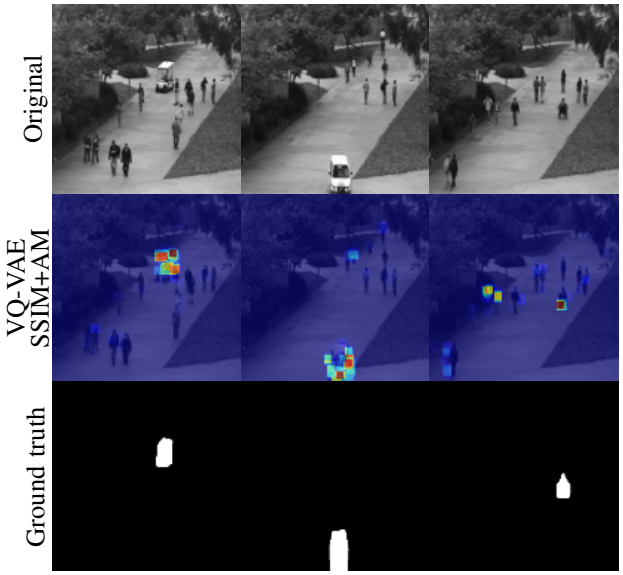


Fig. 3: Selected illustrations of the detection of our model in the UCSD-Ped1 experiment.

D. CIFAR-10 dataset

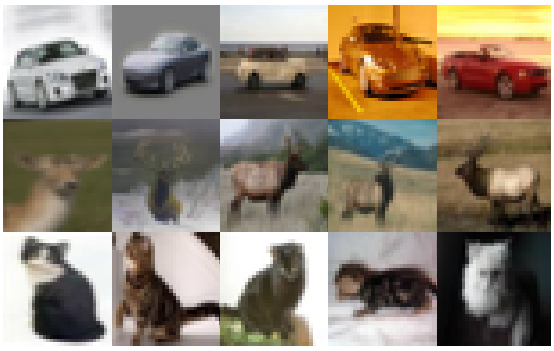
This last experiment addresses the problem of image-wise AD on the CIFAR-10 dataset [22]. The dataset is composed of $60,000$ distributed between $50,000$ train and $10,000$ test images; all equally divided in 10 categories. The original 32×32 size of the images is used without resizing. We test all the models in a *one-versus-rest* way. More precisely, a model is trained on images from one category only and, at test time, we want to discriminate between the *one* class, the normal samples upon which the model was trained, and the *rest* class, the anomalous samples which can come from all the other categories. The metric used is then the image-wise ROCAUC.

We compare our VQ-VAE approaches with a vanilla autoencoder approach [28] and the Deep Support Vector Data Description [29], a standard approach for image-wise AD. To provide an anomaly score for a whole image in the VQ-VAE approaches, we take the average value over all the pixel-wise anomaly map. Note that the VQ-VAE encoder and decoder described in Section III-A are also reduced to two convolutional layers in this experiment. The latent space here has dimension 8×8.

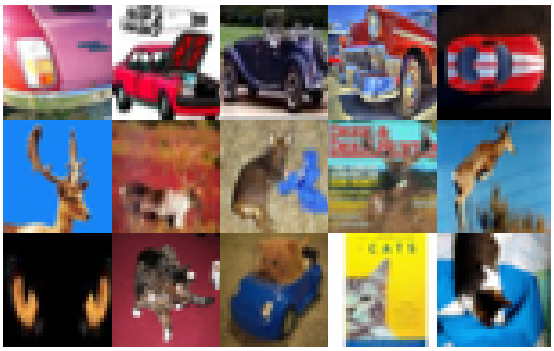
Table III gives the scores for the models over all the categories of the dataset. Scores for the other models are both taken from [29]. We can see that our VQ-VAE approach has a slight advantage against the others which suggests that the provided anomaly map is also relevant for sample-wise AD. Figure 4 illustrates the experiment by highlighting the most normal and anomalous samples according to the VQ-VAE metric. It appears that the VQ-VAE is able to correctly award a high normality score to relatively diverse images inside a same

Category	AE L2 [28]	DSVDD [29]	VQ-VAE SSIM	VQ-VAE SSIM+AM
Airplane	0.59	0.62	**0.71**	0.69
Automobile	0.57	**0.66**	0.63	0.65
Bird	0.49	0.51	0.63	**0.65**
Cat	0.58	0.59	0.62	**0.63**
Deer	0.54	0.61	0.60	**0.64**
Dog	0.62	0.66	**0.67**	**0.67**
Frog	0.51	**0.68**	0.61	0.63
Horse	0.59	**0.67**	0.63	0.63
Ship	**0.77**	0.76	0.74	0.74
Truck	0.67	**0.73**	0.64	0.65
Mean	0.59	0.65	0.65	**0.66**

TABLE III: ROCAUC scores on the CIFAR-10 dataset.



(a) Most normal samples



(b) Most anomalous samples

Fig. 4: Selected illustrations from the CIFAR-10 experiment for the VQVAE SSIM+AM model: some of the most normal (a) and anomalous (b) samples from the *Automobile* (top), *Deer* (middle) and *Cat* (bottom) categories, when the model is trained on this category.

class (colors, scenery, background, etc.). This might reflect that the model extracts relevant features from the images.

IV. Conclusion

In this paper, we showed the potential of VQ-VAEs for AD. We highlighted for the first time that the inner metrics of VQ-VAE models are robust and efficient to detect anomalies in images. Indeed, after developing an intuitive approach to construct an anomaly map, we reported results competitive with several other state-of-the-art approaches for pixel-wise AD on the MVTec and UCSD-Ped1 datasets, as well as for image-wise AD on the CIFAR-10 dataset.

The results show that the inner metrics of VQ-VAEs outperform the inner metrics of VAEs, without introducing more complexity in the model. Our results also corroborate the increasing interest of the machine learning community for VQ-VAEs. In the near future, further study might be conducted to assess whether VQ-VAEs should replace VAEs as the standard architectures in AD workflows.

Acknowledgments

This work was done as a part of the Game of Trawls project. We thank the European Maritime and Fisheries Fund (contract number 18/2216442) and France Filière Pêche (contract number 19/1000544) for funding. It was also supported by the SEMMACAPE project, which benefits from an ADEME (*Agence de la transition écologique*) grant under the "Sustainable Energies" call for research projects (2018–2019).

References

[1] Marco A F Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[2] Douglas M Hawkins, *Identification of Outliers*, Monographs on applied probability and statistics. Chapman and Hall, 1980.

[3] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, 2021.

[4] Lu Wang, Dongkai Zhang, Jiahao Guo, and Yuexing Han, "Image anomaly detection using normal data only by latent space resampling," *Applied Sciences*, vol. 10, no. 23, pp. 8660, 2020.

[5] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft, "Deep semi-supervised anomaly detection," in *8th International Conference on Learning Representations, ICLR*, 2020.

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

[7] Eric T Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan, "Detecting out-of-distribution inputs to deep generative models using a test for typicality," in *4th Workshop on Bayesian deep learning (NIPS)*, 2019.

[8] Sergio Naval Marimont and Giacomo Tarroni, "Anomaly detection through latent space restoration using vector quantized variational autoencoders," in *18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1764–1767.

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

[10] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 289–297.

[11] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," *IEEE transactions on pattern analysis and machine intelligence*, 2021, In press.

[12] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014.

[13] Diederik P Kingma and Max Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[14] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps, "Towards visually explaining variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8642–8651.

[15] David Dehaene, Oriel Frigo, Sébastien Combrexelle, and Pierre Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," in *8th International Conference on Learning Representations, ICLR*, 2020.

[16] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis, "Attention guided anomaly localization in images," in *European Conference on Computer Vision*. Springer, 2020, pp. 485–503.

[17] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[18] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," in *Proceedings of the 38th International Conference on Machine Learning*. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831, PMLR.

[19] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso, "Unsupervised brain anomaly detection and segmentation with transformers," in *Medical Imaging with Deep Learning*. 2021, vol. 143 of *Proceedings of Machine Learning Research*, pp. 596–617, PMLR.

[20] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

[21] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[22] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[23] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Advances in neural information processing systems*, 2019, pp. 14866–14876.

[24] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*. 2019, pp. 372–380, SciTePress.

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] Gabriel Loaiza-Ganem and John P Cunningham, "The continuous Bernoulli: fixing a pervasive error in variational autoencoders," in *Advances in Neural Information Processing Systems*. 2019, vol. 32, Curran Associates, Inc.

[27] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller, "Explainable deep one-class classification," in *9th International Conference on Learning Representations, ICLR*, 2021.

[28] Alireza Makhzani and Brendan J Frey, "Winner-take-all autoencoders," in *Advances in Neural Information Processing Systems*. 2015, vol. 28, Curran Associates, Inc.

[29] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft, "Deep one-class classification," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, PMLR.