remote sensing

MDPI

Communication

A One-Class Classifier for the Detection of GAN Manipulated Multi-Spectral Satellite Images

Lydia Abady *, Giovanna Maria Dimitri and Mauro Barni

Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy; giovanna.dimitri@unisi.it (G.M.D.); barni@dii.unisi.it (M.B.)
* Correspondence: lydia.abady@unisi.it

Abstract: The current image generative models have achieved a remarkably realistic image quality, offering numerous academic and industrial applications. However, to ensure these models are used for benign purposes, it is essential to develop tools that definitively detect whether an image has been synthetically generated. Consequently, several detectors with excellent performance in computer vision applications have been developed. However, these detectors cannot be directly applied as they areto multi-spectral satellite images, necessitating the training of new models. While two-class classifiers generally achieve high detection accuracies, they struggle to generalize to image domains and generative architectures different from those encountered during training. In this paper, we propose a one-class classifier based on Vector Quantized Variational Autoencoder 2 (VQ-VAE 2) features to overcome the limitations of two-class classifiers. We start by highlighting the generalization problem faced by binary classifiers. This was demonstrated by training and testing an EfficientNet-B4 architecture on multiple multi-spectral datasets. We then illustrate that the VQ-VAE 2-based classifier, which was trained exclusively on pristine images, could detect images from different domains and generated by architectures not encountered during training. Finally, we conducted a head-to-head comparison between the two classifiers on the same generated datasets, emphasizing the superior generalization capabilities of the VQ-VAE 2-based detector, wherewe obtained a probability of detection at a 0.05 false alarm rate of 1 for the blue and red channels when using the VQ-VAE 2-based detector, and 0.72 when we used the EfficientNet-B4 classifier.

Keywords: generative adversarial networks; variational autoencoder; EfficientNet; detection; Sentinel-2; remote sensing

Citation: Abady, L.; Dimitri, G.M.; Barni, M. A One-Class Classifier for the Detection of GAN Manipulated Multi-Spectral Satellite Images. *Remote Sens.* **2024**, *16*, 781. https://doi.org/10.3390/rs16050781

Academic Editors: Lionel Bombrun, Keshav D Singh, Sajid Saleem and Abdul Bais

Received: 15 December 2023
Revised: 14 February 2024
Accepted: 21 February 2024
Published: 24 February 2024

1. Introduction

Deep Learning (DL) techniques have established new State of the Art (SOTA) benchmarks in several fields: from bioinformatics to computer vision, from natural language processing to object detection [1–5]. In this context, the development of tools for the creation of image forgeries, on one hand, and for authenticity verification and other forensic applications, on the other, has seen a steep increase in the use of DL techniques [6]. Among the possible application domains, great attention is increasingly being devoted to satellite image analysis.

Satellite images play a crucial role in various application areas, such as meteorological forecasts, landscape analysis, agriculture, regional planning, the monitoring and detection of natural disasters, and several others. As a result, the number of commercial satellites is constantly growing, and the accessibility of satellite images with larger and larger ground resolution [7] is increasing daily. As for other application domains, DL provides various tools for manipulating satellite images. Some examples of DL-based tools for satellite image manipulations are described in Refs. [8–10]. Such manipulations are often related to disinformation campaigns, as reported, for instance, in [11]. Hence, there is a growing need to develop DL forensic methods suited for the detection and identification of satellite

image forgeries. The extension of image forensics tools developed for computer vision applications to satellite imagery, however, has proven to be challenging from several points of view. First of all, forensic techniques developed for non-satellite images must be adapted to the specific content of overhead imagery. This is due to the inherently different types of features and characteristics of conventional RGB and multi-spectral satellite images. In addition, quite often, satellite images have more than three bands. For example, Sentinel-2 Level 1C optical images have 13 bands, with each group of bands characterized by a different ground sampling distance. Moreover, unlike RGB images, each pixel is typically represented by more than 8 bits per band (12 bits in the case of Sentinel-2 Level 1C images). On top of that, synthetically generated datasets of multi-spectral images are missing, thus making it difficult to benchmark DL tools for satellite image forgery detection.

Over the years, Generative Adversarial Networks (GAN) have been extensively researched and applied to various domains, including image generation [12,13], style transfer [14,15], super-resolution [16], text-to-image synthesis [17,18], and more. Moreover, researchers have proposed numerous variants and improvements to the original GAN architecture, such as conditional GANs [19], Wasserstein GANs [20], cycle-consistent GANs [21], and progressive growing GANs [22], among others. GANs have also been used for unsupervised representation learning [23], where the generator learns to generate data samples that capture the underlying data distribution, enabling downstream tasks such as classification and clustering to be performed with improved accuracy. Additionally, GANs have found applications beyond image generation, including in the fields of natural language processing (NLP) [24], speech synthesis [25], and generative music composition [26].

GANs can be successfully used to create synthetic satellite images [8]. Such architectures have proven to be extremely useful for image generation or style transfer, and they apply not only to RGB images [12,19] but also to multi-spectral images [27], with only some minor modifications. A few works have also been proposed for the detection of GAN-generated satellite images. In general, most of these tools are only trained on the RGB bands of multi-spectral images and exhibit good detection capabilities when the training and test datasets are acquired under matched conditions, but they fail to generalize to unseen data. One tool that is actually trained on multi-spectral images but also lacks generalization capabilities is described in [28], where the authors present a detector based on an EfficientNet-B4 [29] model trained on multi-spectral images, achieving very good performance. When the model was tested on a different dataset, however, the performance declined significantly.

To overcome the generalization weakness of SOTA techniques for the detection of GAN multi-spectral satellite images, in this work, we propose using a one-class classifier based on a Vector Quantized Variational Autoencoder (VQ-VAE) trained only on pristine images. Moreover, we propose using the reconstruction loss between the input and the output images to distinguish GAN and pristine images.

We evaluated the performance of the one-class classifier on several multi-spectral GAN synthetic Sentinel-2 level-1C satellite datasets and compared them against those of a conventional manipulation detector based on EfficientNet-B4, which was trained on both pristine and GAN-generated satellite images. We ran experiments on the full 13 bands of Sentinel-2 level-1C samples. The results we obtained demonstrated the superior performance of the one-class classifier in terms of generalization capability, with only a small performance loss compared to the two-class classifier under matched conditions.

The paper is structured as follows: in Section 2, we overview the related works in the field of satellite imagery forgery detection. In Section 3, we describe the datasets we created or collected, differentiating between the various models and GANs methodologies used. In Section 4, we describe the VQ-VAE one-class classifier, and in Section 5, we present the experimental results proving the validity of the proposed method. Finally, Section 6 concludes and offers future perspectives on our work.

2. Related Work

In this section, we overview the prior work on satellite image forgery detection. In [30], the authors introduced a framework composed of two steps for the detection and localization of forgeries in satellite images. In the first step, a GAN is trained to obtain a set of features capable of representing pristine satellite images. In the second step, a one-class classifier based on Support Vector Machine (SVM) is applied to distinguish pristine and non-pristine images. The method proposed in [31] is based on a conditional GAN architecture that is trained on two domains, the first domain is the domain of spliced images, while the second contains the forgery masks. Given a GAN image as input, a GAN generator is used to estimate a forged mask that is as close as possible to the real one. In [32], the authors applied an analysis similar to [30], by jointly training the auto-encoder and a support vector data descriptor [33]. Furthermore, in [34], the authors took advantage of a one-class deep belief network, which was employed for detecting and localizing forged images. The same authors, in [35], proposed using a vision transformer for image reconstruction. In this case, the forgery mask was obtained by observing the differences between the input and output images.

A further interesting work is [36], where the heat maps of forged regions were estimated using a novel architecture based on a U-Net nested within a GAN architecture. The system built in this way could localize RGB image forgeries generated by three different types of GANs: StyleGAN2 [37], ProGAN [22], and CycleGAN [21]. The forgeries were created using Sentinel-2 RGB images that were spliced within the generated images. The output of the architecture consisted of a probability mask, where each pixel was associated with the probability of having been generated by one of the specific GANs that were tested in the experimental framework.

In [10], the authors implemented a GAN-based approach for semantic satellite image translation. In the same work, they also presented a data-driven approach for the detection of GAN generated images and used SVM to detect cycleGAN-generated images from a set of features (both spatial and spectral). Furthermore, in [38], the authors relied on the dataset presented in [5], to develop an additional method to detect RGB satellite images that are semantically transformed, starting from the detection of high-frequency details in the generated samples.

Most of the methods that have been proposed so far focus on RGB 8-bit images. To the best of our knowledge, the only method that has been proposed for detecting multi-spectral image forgeries of images with 13 bands, like those acquired by Sentinel-2 sensors, is [28], which cannot generalize to mismatched data, as we previously mentioned in the Introduction.

3. Datasets

The datasets we used consist of two main types of images: pristine Sentinel-2 level1-C images and *GAN* images generated from models trained on Sentinel-2 level1-C datasets. The details of acquiring the pristine images and generating synthetic images can be found in the following article [39]. The images consist of 13 bands of 512 $\times$ 512 patches. For the generalization experiments, we collected an additional dataset, hereafter referred to as the "this-city-does-not-exist" dataset, containing RGB images of size 1024 $\times$ 1024. A short description of all the datasets is given in the following subsections. In addition, a summary of all the datasets described in this section is given in Table 1.

The datasets were employed in various tasks. To start with, we needed a set of pristine images to train the VQ-VAE 2-based detector, pristine images, and GAN images to be used to train the 2-class EfficientNet detector, as well as GAN images to be used to test both detectors. A small number of pristine images were used to calibrate the thresholds for both detectors since the assessment metric we used was the probability of detection at a false alarm rate of 0.1. Table 2 shows how we split the images of the various datasets across different tasks. For VQ-VAE 2 training, only pristine images were used to train a single one-class detector. For EfficientNet-B4, the training images were used to train several

versions of the detectors, each time using a different combination of the training sets (see Section 5.1). The images used for testing were never used during training.

Table 1. Summary of the datasets used in our work (see Section 3).

Dataset	Bands	Size	Architecture	Transfer Type	Total # Pristine Source: Sentinel2-level1C	Total # GAN
Land Cover (LC)	13	512 × 512	CycleGAN	Land Cover	30,000	4000
Scandinavian (Scand)	13	512 × 512	Pix2pix	Season	17,044	4000
China	13	512 × 512	Pix2pix	Season	16,000	4000
Alps	13	512 × 512	–	–	7872	0
This-city-does-not-exist	3	1024 × 1024	styleGAN2	–	0	140

Table 2. Summary of the number of images used for each task. P indicates pristine images and G denotes GAN-generated images.

Dataset	Train EfficientNet-B4	Train VQ-VAE 2	Test Detectors	Calibrate Threshold of Detectors
LC	3000 P, 3000 G	29,000 P	1000 G	100 P
Scand	3000 P, 3000 G	–	1000 G	100 P
China	3000 P, 3000 G	15,000 P	1000 G	100 P
Alps	–	7872 P	–	–
This-city-does-not-exist	–	–	140 G	–

3.1. Land Cover (LC) Transfer Datasets

The first dataset was the land cover (LC) transfer dataset. This dataset was formed of two classes of images: pristine and GAN images. As indicated in Table 2, 29,000 pristine images were used to train the VQ-VAE 2 image detector and 1000 images were set aside for testing. For training the EfficientNet-B4 detector, we used 3000 pristine images and 3000 generated images, and tested on 1000 GAN and 1000 pristine images, while 100 pristine images were used for threshold calibration. Figure 1a shows two examples from the pristine LC dataset. Figure 1b, instead, shows two examples of images generated by the cycleGAN.

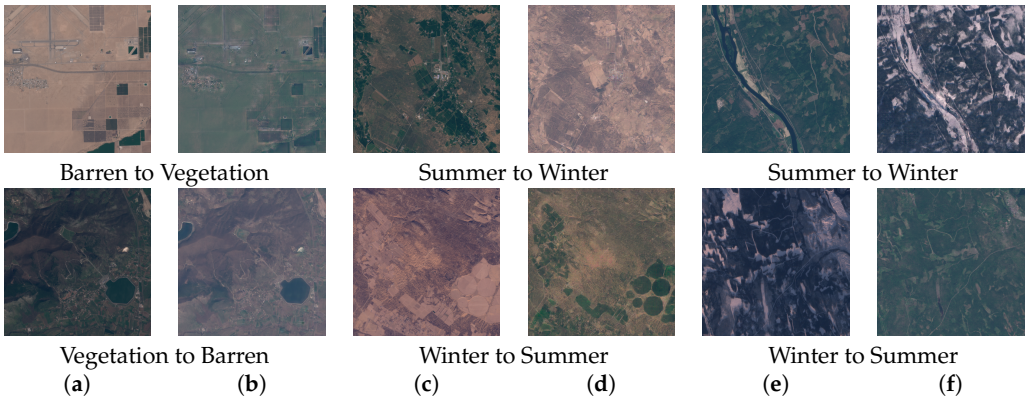


Figure 1. Some examples of the images contained in the datasets used throughout the paper. Only the RGB bands are shown. (**a**) LC-Pristine; (**b**) LC-GAN; (**c**) China-Pristine; (**d**) China-GAN; (**e**) Scand-Pristine; (**f**) Scand-GAN.

3.2. China and Scandinavian Season Transfer Datasets

The China and Scandinavian (Scand) Season Transfer datasets are another example of image-to-image GAN image generation. For each dataset, we used 6000 images to train the EfficientNet-B4 detector, 100 pristine images were used for threshold calibration, and 1000 GAN images were used to test the two detectors. In addition, 15,000 pristine images from the pristine China dataset (excluding the images that were used for calibration and testing) are used as part of the VQ-VAE 2 training dataset. Figure 1c shows two examples of pristine images, while Figure 1d shows two examples of GAN-generated images. Figure 1e,f show two examples of pristine and GAN images.

3.3. Alps Dataset

The *Alps* dataset is a pristine image dataset that was collected to help train the VQ-VAE 2 detector. We collected images from the same area in two different months, with each month representing a different season (June 2019 for summer and December 2019 for winter). To avoid generating images with clouds, we selected images with limited cloud cover. Since it was not possible to obtain images with 0% cloud cover, we limited the search to images with cloud cover less than 9%. As a result, we obtained a dataset with 7872 pristine images. Figure 2a shows an RGB representation of two examples of the Alps dataset.

(**a**) (**b**)

Figure 2. Examples of images taken from this-city-does-not-exit and Alps datasets, where for the Alps dataset only RGB bands are shown. (**a**) Alps dataset. (**b**) This-city-does-not-exist.

3.4. This-City-Does-Not-Exist Dataset

This dataset was used to test the ability of the various models to generalize to images generated from unknown architectures and with completely different contents. It contains only GAN images downloaded from [40]. The images were generated using a styleGAN2 model. We collected 140 images of size 1024×1024. The images of this dataset have only three bands (RGB). Figure 2b, shows two examples of the this-city-does-not-exist dataset.

4. The VQ-VAE 2 One-Class Classifier

In this section, we describe the one-class classifier we developed to distinguish pristine and GAN multi-spectral images. We start with a brief introduction to autoencoders and variational autoencoders, then we describe the VQ-VAE 2 architecture, which is the one our system relies on.

A neural network A that is trained to reconstruct its input at the output is referred to as an autoencoder [41]. The reconstruction is constrained in such a way as to prevent learning the identity function. An autoencoder is divided into two main parts:

- The encoder A_{e}, mapping the input x onto a hidden representation h (i.e., $h = A_{\mathrm{e}}(x)$).
- The decoder A_{d}, reconstructing an approximate version of the input $\tilde{x}$ from the hidden representation (i.e., $\tilde{x} = A_{\mathrm{d}}(h)$).

In the case of tensor data, the input can be an image x and the hidden representation a vector h. The encoder and decoder are trained jointly to reduce the reconstruction loss, usually a L_2 loss term, between the input and output samples.

In Variational Autoencoders (VAEs) [42], the input is encoded into a vectorial representation, and the hidden representation's features are forced to follow a Gaussian distribution that is denoted by $\mathcal{N}(f(x); g(x))$, where $f(\cdot)$ denotes the mean and $g(\cdot)$ denotes the variance of the distribution. A sample of the hidden representation is taken during the decoding stage and is utilized as input to the decoder, which produces a reconstructed version of the original input data. When the hidden features are required to follow a Gaussian distribution, the total loss that is used during training is equal to

$$L(x, \tilde{x}) = \|x - \tilde{x}\|_2^2 + \beta L_{KL}(\mathcal{N}(f(x), g(x)), \mathcal{N}(0, I_d)), \quad (1)$$

where the first term is a "data fidelity term" that measures the difference between the input sample x and the estimated sample $\tilde{x}$, the second term applies a kind of "regularization" by requiring the network to minimize the Kullback–Leibler divergence L_{KL} between the learned hidden variable distribution and a desired normal distribution $\mathcal{N}(0; I_d)$ (here, I_d is the identity matrix), and β is a hyperparameter balancing the two loss terms. The decoder is used to create new images by selecting random samples from the hidden layer after training the VAE.

VQ-VAE [43] is a variant of VAE that uses vector quantization to learn a discrete latent representation instead of a continuous one. This is done by adding a discrete codebook component to the network that contains the list of vectors with their indices. Then, the output of the encoder is compared with all the codebook entries in terms of Euclidean distance, and the code that is closer to the output of the encoder is fed to the decoder. In VQ-VAE, compared to a VAE, the priors are learned rather than taken as static input. The combination of a discrete latent representation and an autoregressive prior thus paves the way for the generation of high-quality images, videos, and speech.

VQ-VAE 2 [44] is similar to VQ-VAE, with the only difference being that it uses multi-scale latent maps to increase the resolution of the reconstructed image. In our experiments, we used three levels of latent maps and only relied on a static prior instead of an autoregressive prior, since our goal was to detect GAN images rather than generate them. Figure 3 shows the architecture we used, where we opted for a three-level hierarchy: bottom, middle, and top, with latent space sizes of 512, 128, and 64, respectively. The input to the architecture is one of the bands of the pristine image and the output is the reconstructed version of the input.

Concerning the detection of GAN images, we used the VQ-VAE 2 architecture according to two different modalities. In the first approach, the autoencoder processes all 13 bands together (the resulting architecture is referred to as VQ-VAE 2_{13}. For the second approach, we trained one model per band (referred to as VQ-VAE 2_1). The reconstruction loss on all bands and the total reconstruction loss on all bands were used as features to be processed by an anomaly detection module, e.g., a one-class SVM. Based on the experiments we carried out (see Section 5.2), we decided to use the reconstruction loss directly and detect the GAN images by applying a threshold to the reconstruction error band-by-band.

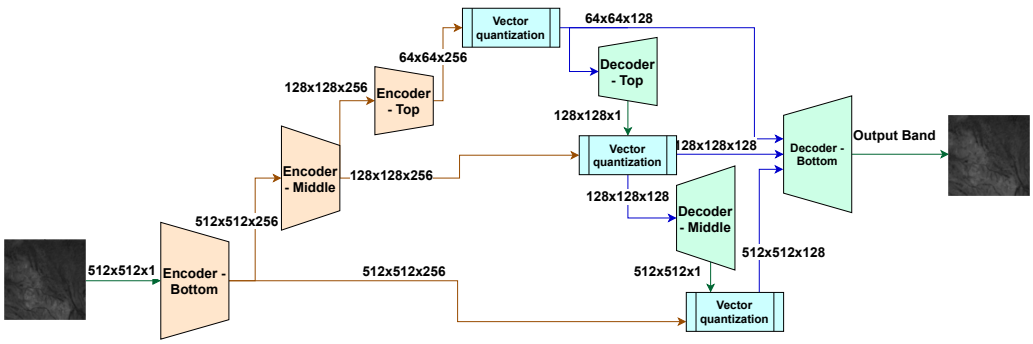


Figure 3. VQ-VAE 2 Architecture.

5. Experiments and Results

In this section, we evaluate the performance of the proposed VQ-VAE 2 and empirically prove that an off-the-shelf 2-class detector based on EfficientNet-B4 lags behind the VQ-VAE 2 in terms of generalization capabilities. Both detectors were trained and tested with the datasets described in Section 3.

5.1. EfficientNet-B4 Detector

As a baseline 2-class detector to benchmark the performance of our system, we trained several models based on the EfficientNet architecture. The EfficientNet class of networks was proposed as a way to efficiently scale the network depth, width, and resolution based on the input dimensions [29]. In our experiments, we used EfficientNet-B4 (eff_down), with hyper-parameters set as in [29], the only exception being the difference in the input size, which we adapted to match the number of channels our images consisted of (13). We built four different models by training the networks on the LC dataset (thus fitting to cycleGAN data), the Scandinavian dataset (adapting the model to distinguish pix2pix images), the China dataset, and a combination of the LC and Scandinavian datasets. The four models were cross-tested on the LC, Scandinavian, and China datasets. We also trained three additional models, this time by removing the downsampling from the initial layer, as suggested by [45], to enhance the generalization capabilities of the model. We called these models EfficientNet-B4 with no down (eff_nodown). Augmentation, including Gaussian blur, random shift, random rotation, and random flip, was applied to train all models.

The results we obtained on the various datasets are shown in Table 3, reporting the correct detection probability at a false alarm rate equal to 0.1. For threshold calibration, for each test, we used 100 pristine images from the dataset to be tested. We also obtained other results that we do npt report here using different thresholds, where we obtained the thresholds from 100 pristine images from the corresponding training dataset. However, the conclusions drawn were the same.

As expected, the probability of detection was very good when the datasets used for training and testing were matched; that is, when the GAN images were generated by the same GAN model used for training. Concerning generalization, we observed that the eff_nodown architecture had better generalization capabilities. In any case, the performance dropped when the models were tested on images taken from datasets that were not used during training. The best results were obtained by training the detector on images generated by both pix2pix and cycleGAN. Even in this case, however, the performance deteriorated when the detector was tested on the images of the China dataset that had not been used during training.

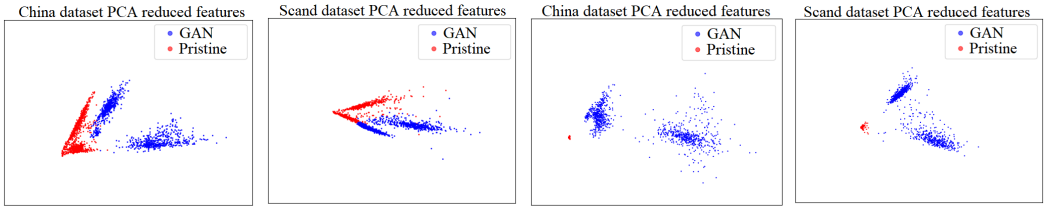
Table 3. EfficientNet-B4 correct detection probability (Pd) (false alarm rate (FAR) was set to 0.1).

Pd@0.1 FAR			**Test**		
			LC	**Scand**	**China**
Train	**LC**	**eff_down**	1	0.8	0.7
		eff_nodown	1	1	0.82
	Scand	**eff_down**	0.61	1	0.65
		eff_nodown	0.73	1	0.74
	China	**eff_down**	0.66	0.68	1
		eff_nodown	0.71	0.75	1
	LC & Scand	**eff_down**	1	1	0.6
		eff_nodown	1	1	0.86

5.2. Vector Quantized Variational Autoencoder 2

To build the VQ-VAE 2 one-class classifiers, we trained a VQ-VAE 2_{13} model working on all bands together, and 13 VQ-VAE 2_1 models, each working on one different band. The models were trained on 50000 pristine Sentinel 2 level-1C images collected from the Alps, China, and land cover datasets. The networks were trained for 100 epochs, with early stopping and a batch size equal to 64.

Some initial insights into the discrimination capability of the VQ-VAE 2 models could be obtained by plotting the reconstruction losses obtained, applying the trained autoencoders to the various datasets. To obtain the scatter plots shown in Figure 4, we applied principle component analysis (PCA) to reduce the feature dimensionality to 2. The scatter plots obtained using the models trained on VQ-VAE 2_{13} and VQ-VAE 2_1 are reported for two datasets: the China dataset, whose pristine images are part of the VQ-VAE 2 training dataset, and the Scandinavian dataset, that was never seen by the VQ-VAE 2. We can observe that, in both cases, pristine and GAN images are grouped into well-distinct clusters; however, the features are further apart in the case where the reconstruction losses were taken from the 13 models of VQ-VAE 2_1. Hence, the rest of our experiments were carried out by training one model per band.



(**a**) (**b**)

Figure 4. Scatter plot after PCA feature reduction for GAN and pristine image datasets. (**a**) VQ-VAE 2 trained on 13 bands. (**b**) VQ-VAE 2 trained band by band.

Given that 13 single-band trained autoencoders, we used the reconstruction error of each autoencoder to detect the GAN images. To do so, we set the detection threshold by fixing the false alarm rate on 100 pristine images for each testing dataset to 0.1. Similarly to the EfficientNet-B4 results, we also experimented with obtaining the threshold from the training dataset, which in this experiment was obtained from the China and LC datasets. The results, we obtained were slightly worse for the 10 m bands but were comparable for the rest of the bands.

The tests were carried out on the LC, Scand, and China datasets, as well as on all the datasets mixed together. Table 4 shows the results we obtained. The correct detection probability was very good for most of the bands, with some room for improvement for the R (band 4), G (band 3), B (band 2), and NIR (band 8) bands. The results in Table 4 show the excellent generalization capability of the VQ-VAE 2-based detectors, as they could detect the GAN images even in the Scand dataset (whose pristine images were not part of the training dataset) by looking at any of the bands except bands 2 and 3. In particular, the detectors based on bands 7, 8a, 9, 11, and 12 achieved nearly perfect results for all datasets.

Table 4. Correct detection probability at a 0.1 false alarm rate using VQ-VAE 2 for the 13-band datasets.

BANDS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**8a**	**9**	**10**	**11**	**12**
LC	0.94	0.88	0.9	0.55	0.83	0.92	0.99	0.7	0.99	0.98	0.99	1	1
Scand	0.98	0.64	0.51	0.95	1	0.97	0.98	0.97	0.99	1	0.99	0.97	0.99
China	0.99	0.98	0.99	0.98	1	0.98	0.99	0.73	0.99	0.99	0.54	1	1
Mixed data	0.85	0.86	0.6	0.81	0.92	0.95	0.98	0.7	0.99	0.98	0.85	1	1

5.3. Comparison on the City-Does-Not-Exist Dataset

To further test the generalization capabilities of the various detectors, we used the "this-city-does-not-exist" dataset. The reasons for this choice were twofold. The first reason was that these images were generated by styleGAN 2, which is an architecture that was not used in our experiments. The second reason was that the VQ-VAE 2 autoencoder was trained on the same pristine images as the GAN architectures, while for the "this-city-does-not-exist" dataset, the GAN training dataset was unknown to us (and surely different from that we used to train the autoencoder). To carry out these experiments, we trained an EfficientNet-B4 no down detector on the LC and Scand datasets, but using only the 3 RGB channels. As for the VQ-VAE 2, we used the same single-band models that had been trained before (of course in this case we had to use only the detectors working on the R, G, and B bands). The detection thresholds of all detectors were fixed by using a set of pristine Sentinel-2 images, the same 100 we had used before in the case of the LC and Scand datasets, this time targeting a false alarm rate equal to 0.05.

The results of these experiments are shown in Table 5. As we can see, the VQ-VAE 2 detector provided much better results than EfficientNet-B4, which was not able to properly detect the images generated from scratch by the styleGAN 2 generator. This was not the case with the VQ-VAE 2 detector, which could detect the styleGAN 2 images without retraining.

Table 5. Correct detection probability at a 0.05 false alarm rate on the "this-city-does-not-exist" dataset.

Metrics	VQ-VAE 2 (Red)	VQ-VAE 2 (Blue)	VQ-VAE 2 (Green)	EfficientNetB4 (Trained on 3 Bands)
Pd@0.05 FAR	1	1	0.96	0.72

6. Conclusions

We introduced a one-class detector of GAN multi-spectral images generated by a variety of DL architectures. The model is based on a VQ-VAE 2 autoencoder and was trained only on pristine images. To the best of our knowledge, this is the first work proposing the use of a one-class classifier to detect 13-band Sentinel-2 level-1C artificially generated images.

We ran experiments on images that had been generated by cycleGAN and pix2pix architectures. The results we obtained are particularly promising. In particular, the proposed detector exhibited a superior generalization capability than a baseline 2-class detector based on EfficientNet-B4. To evaluate the generalization capability in extreme conditions, we tested the detector on a small dataset of RGB satellite images generated by styleGAN 2. The proposed detector outperformed the two-class classifier by a wide margin. Further work could be performed to diversify the generative models that were used to create the satellite images, to include additional GAN and diffusion models.

Author Contributions: Conceptualization, L.A., and M.B.; methodology, L.A.; software, L.A.; validation, L.A.; formal analysis, L.A., and M.B.; investigation, L.A.; data curation, L.A.; writing—original draft preparation, L.A.; writing—review and editing, M.B. and G.M.D.; visualization, L.A.; supervision, M.B. and G.M.D.; project administration, M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) grant number FA8750-20-2-1004.

Data Availability Statement: Data is available on request, code is available: vqvae2: https://github.com/lydialy8/vqvae2_based_classifier (accessed on 24 January 2024) efficientnetb4: https://github.com/lydialy8/eff_nodown_rs (accessed on 24 January 2024)

.

Acknowledgments: This article is funded by the DARPA and AFRL. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or AFRL or the U.S. Government.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Briefings Bioinform.* **2017**, *18*, 851–869. [CrossRef] [PubMed]
2. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
3. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing.*IEEE Trans. Neural Networks Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
4. Dimitri, G.M.; Spasov, S.; Duggento, A.; Passamonti, L.; Lió, P.; Toschi, N. Multimodal and multicontrast image fusion via deep generative models. *Inf. Fusion* **2022**, *88*, 146–160. [CrossRef]
5. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Networks Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
6. Yang, P.; Baracchi, D.; Ni, R.; Zhao, Y.; Argenti, F.; Piva, A. A survey of deep learning-based source image forensics. *J. Imaging* **2020**, *6*, 9. [CrossRef] [PubMed]
7. High Resolution Satellite Data. Available online: https://landinfo.com/worldwide-mapping-products/high-resolution-global-satellite-imagery/ (accessed on 22 November 2022).
8. Abady, L.; Horváth, J.; Tondi, B.; Delp, E.J.; Barni, M. Manipulation and generation of synthetic satellite images using deep learning models. *J. Appl. Remote Sens.* **2022**, *16*, 046504. [CrossRef]
9. Baier, G.; Deschemps, A.; Schmitt, M.; Yokoya, N. Synthesizing Optical and SAR Imagery From Land Cover Maps and Auxiliary Raster Data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
10. Zhao, B.; Zhang, S.; Xu, C.; Sun, Y.; Deng, C. Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartogr. Geogr. Inf. Sci.* **2021**, *48*, 338–352. [CrossRef]
11. Australian Misleading Fires News. Available online: https://www.bbc.com/news/blogs-trending-51020564 (accessed on 19 November 2022).
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
13. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410. [CrossRef]
14. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423. [CrossRef]
15. Huang, X.; Belongie, S.J. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 1510–1519. [CrossRef]
16. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.P.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 105–114. [CrossRef]
17. Zhang, H.; Xu, T.; Li, H. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 5908–5916. [CrossRef]
18. Jiang, B.; Huang, Y.; Huang, W.; Yang, C.; Xu, F. Multi-scale dual-modal generative adversarial networks for text-to-image synthesis. *Multimed. Tools Appl.* **2023**, *82*, 15061–15077. [CrossRef]
19. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
20. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
21. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2242–2251. [CrossRef]
22. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.

23. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York, NY, USA, 19–24 June 2016*; Balcan, M., Weinberger, K.Q., Eds.; JMLR Workshop and Conference Proceedings; JMLR: Brookline, MA, USA 2016; Volume 48, pp. 1558–1566.
24. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017*; Singh, S., Markovitch, S., Eds.; AAAI Press: Washington, DC, USA, 2017; pp. 2852–2858. [CrossRef]
25. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.Ö.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
26. Yang, L.; Chou, S.; Yang, Y. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, 23–27 October 2017*; Cunningham, S.J., Duan, Z., Hu, X., Turnbull, D., Eds.; Ubiquity Press: London, UK, 2017 ; pp. 324–331.
27. Abady, L.; Barni, M.; Garzelli, A.; Tondi, B. GAN generation of synthetic multispectral satellite images. In Proceedings of the Image and Signal Processing for Remote Sensing XXVI, International Society for Optics and Photonics, Online, 21–25 September 2020; Volume 11533, pp. 122–133.
28. Abady, L.; Dimitri, G.M.; Barni, M. Detection and Localization of GAN Manipulated Multi-spectral Satellite Images. In Proceedings of the 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 5–7 October 2022; pp. 339–344.
29. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
30. Yarlagadda, S.; Güera, D.; Bestagini, P.; Zhu, F.; Tubaro, S.; Delp, E. Satellite image forgery detection and localization using GAN and one-class classifier. In Proceedings of the Electronic Imaging (EI), San Francisco, CA, USA, 28 January–1 February 2018. [CrossRef]
31. Bartusiak, E.R.; Yarlagadda, S.K.; Güera, D.; Bestagini, P.; Tubaro, S.; Zhu, F.M.; Delp, E.J. Splicing detection and localization in satellite imagery using conditional GANs. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019. [CrossRef]
32. Horvàth, J.; Güera, D.; Yarlagadda, S.K.; Bestagini, P.; Zhu, F.M.; Tubaro, S.; Delp, E.J. Anomaly-based manipulation detection in satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), Long Beach, CA, USA, 16–17 June 2019.
33. Tax, D.M.J.; Duin, R.P.W. Support Vector Data Description. *Mach. Learn.* **2004**, *54*, 45–66. [CrossRef]
34. Horvàth, J.; Mas Montserrat, D.; Hao, H.; Delp, E.J. Manipulation detection in satellite images using deep belief networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), Seattle, WA, USA, 14–19 June 2020.
35. Horváth, J.; Baireddy, S.; Hao, H.; Montserrat, D.M.; Delp, E.J. Manipulation Detection in Satellite Images Using Vision Transformer. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021; pp. 1032–1041. [CrossRef]
36. Horváth, J.; Montserrat, D.M.; Delp, E.J.; Horváth, J. Nested Attention U-Net: A Splicing Detection Method for Satellite Images. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges, Virtual, 10–15 January 2021; pp. 516–529.
37. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 8107–8116. [CrossRef]
38. Chen, H.S.; Zhang, K.; Hu, S.; You, S.; Kuo, C.C.J. Geo-DefakeHop: High-Performance Geographic Fake Image Detection. *arXiv* **2021**, arXiv:2110.09795.
39. Abady, L.; Barni, M.; Garzelli, A.; Tondi, B. Generation of synthetic generative adversarial network-based multispectral satellite images with improved sharpness. *J. Appl. Remote Sens.* **2024**, *18*, 014510. [CrossRef]
40. City Does Not Exist. Available online: https://thiscitydoesnotexist.com/ (accessed on 30 March 2022).
41. Hinton, G.E.; Zemel, R. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Denver, CO, USA, 30 November–3 December 1993.
42. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *arXiv* **2019**, arXiv:abs/1906.02691.
43. van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6309–6318.

44. Razavi, A.; van den Oord, A.; Vinyals, O., Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates Inc.: Red Hook, NY, USA, 2019.
45. Gragnaniello, D.; Cozzolino, D.; Marra, F.; Poggi, G.; Verdoliva, L. Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo, ICME 2021, Shenzhen, China, 5–9 July 2021; pp. 1–6.