Anomaly Detection in Satellite Imagery

1st Raúl Barba Rojas
Master Thesis
Universidad Internacional Menéndez Pelayo
barbarojasraul@gmail.com

***Abstract*—Satellite imagery provides society with data to solve a wide range of problems, including monitoring environmental changes, identifying illicit activities, security surveillance and detecting objects of interest, among many others. Anomaly detection has gained attention in recent years, as it can be used to address the aforementioned problems effectively. Traditional anomaly detection methods pose challenges, as they rely on labelled datasets, which are often scarce, expensive to obtain, and limited in capturing the full spectrum of real-world anomalies. To address these challenges, self-supervised learning has emerged as a powerful paradigm that leverages the usage of anomaly detection techniques on unlabelled data, allowing for cheaper solutions that can learn robust representations of known and unknown anomalies in many fields, including earth observation-related fields.**

In this work, we explore state-of-the-art self-supervised anomaly detection methods and evaluate them on two well-established satellite imagery datasets: Landcover.ai, which focuses on land cover classification, and HRC-WHU, which provides high-resolution cloud cover annotations. These datasets allow for a detailed assessment of the effectiveness of these methods in detecting anomalies across diverse geospatial contexts. Furthermore, we propose a novel ensemble self-supervised learning approach that integrates a high- and low-level of abstraction level of abstraction learning objectives to enhance anomaly detection performance. Our method achieves promising results, surpassing existing self-supervised techniques in both datasets.

***Index Terms*—anomaly detection, deep learning, computer vision, self-supervised learning, earth observation, satellite imagery**

I. Introduction

Earth Observation (*EO*) is an active field of research focused on utilising information generated by observing systems, e.g. satellites and unmanned vehicles, to measure multiple aspects of the Earth system, including agricultural monitoring, weather pattern detection and maritime security, among many others [1].

Among the various technologies used for EO, satellite imagery stands out as a fundamental tool, offering high-resolution and real-time, or pseudo-real-time, data for monitoring environmental, geological, and human activities. Satellite imagery has already been used in combination with machine learning techniques, mainly deep learning techniques, in various fields, such as sustainable development [2], maritime security [3], active fire detection [4], and smart agriculture [5].

While satellite imagery can be used under different perspectives to solve the stated problems, anomaly detection has gained attention in recent years. Anomaly detection refers to the problem of detecting abnormal, i.e. unexpected or out of the ordinary, behaviours or patterns in data [6]. Anomaly detection techniques have gained popularity over the past years because they overcome one of the main limitations in satellite imagery studies: the lack of quality labelled data [7], [8]. Labelled data has become a limitation for these studies for multiple reasons: (I) there may be a limited number of labelled datasets that can be used in the domain of interest; (II) there may not be labelled datasets in the domain of interest and the cost of labelling data is always high, as it requires hours of manual work, which can also introduce errors; and (III) anomalies may not always be known, i.e., there exist domains where normal behaviours or patterns are known while abnormal patterns remain unknown [9]. These limitations have increased the popularity of self-supervised learning techniques (*SSL*) anomaly detection techniques in satellite imagery, and many works have been published aiming to provide effective solutions to the problem of detecting anomalies without the need of labelled data [10], [11].

This study has a two-fold objective. On the one hand, we aim to evaluate the effectiveness of existing SSL anomaly detection methods in the field of satellite imagery, understanding their advantages, disadvantages and limitations. Furthermore, we propose a novel method, *DualAnoDAE*, which covers some of the limitations of the existing methods and achieves – arguably surpasses – state-of-the-art results. As reproducibility is an existing concern in the scientific community, all the code developed to carry out this study has been made publicly available at this GitHub repository, including instructions to reproduce the obtained results. The datasets selected for the study, described in the following sections, are publicly available datasets, thus the study is completely reproducible.

The rest of this document is organised as follows: Section II describes classical and state-of-the-art approaches to anomaly detection and their applications in satellite imagery. Section III describes the procedures, techniques and datasets utilised in this study, whereas Section IV describes the results obtained. Section V describes the main limitations in the conduction of this work, and Section VI provides a brief summary of the work, its contributions and future work.

II. Background

Anomaly detection has been studied thoroughly since the creation of the very first techniques. Due to the limitations and challenges posed by the lack of labelled data and the need to detect previously unseen anomalies, self-supervised learning methods trained on ordinary data, i.e. non-anomalous,

are commonly used to solve anomaly detection problems in its various fields of application.

However, the most traditional approaches to anomaly detection, even in satellite imagery, were based on manually extracting visual features from known, non-anomalous, data, which presented non-trivial challenges, such as the curse of dimensionality and the extraction of non-optimal features from the data [12], [13]. In the literature, there is a vast number of works that propose and evaluate multiple traditional anomaly detection methods, including: K-Nearest Neighbors (*KNN*), a distance-based method that calculates the neighbors of each data point and uses them to compute the anomaly score for each node [14]; Local Outlier Factor (*LOF*), a technique derived from KNN based on computing the local density of each point to calculate its anomaly score [15]; Isolation Forests, a traditional method that isolates the anomalous data, characterised by significantly different values in one or more attributes, using a binary tree structure [16]; or One-Class Support Vector Machine (*OCSVM*), which utilises the classical machine learning algorithm known as support vector machine (*SVM*) to learn the normal samples of a given dataset, so that it could be used to detect the non-anomalous samples [17]; among many other approaches [12].

To overcome the limitations imposed by manual feature extraction and the curse of dimensionality inherent to high-resolution satellite imagery, multiple methods were created under the paradigm of Self-Supervised Deep Learning for anomaly detection in various fields. In all these cases, the general methodology employed for anomaly detection in satellite imagery is similar: (I) Firstly, models are trained on the non-anomalous portion of data, which leads to networks, in this case AEs, that learn the representation of the non-anomalous data; (II) Secondly, an anomaly score function is defined. This function is used to determine whether a given sample is anomalous or not; (III) Thirdly, to determine whether a sample is anomalous, an anomaly score threshold must be defined and used. Typically, the anomaly score threshold is computed as the average value of the anomaly score function in the validation data, which must not contain anomalies in the case of SSL-based anomaly detection methods. However, other approaches have also been tried, including the median value and the usage of the IQR in the anomaly score distribution of the validation data [18], [19].

One of the earliest works in neural network SSL-based anomaly detection methods is the AutoEncoder (*AE*), a deep neural network that consists of two main components: the encoder and the decoder. The encoder is the part of the network that transforms the input tensor into an embedded representation of it, which serves as the input of the decoder network, whose purpose is to transform the embedding into an output tensor that must be as similar to the input tensor as possible [20]. This method has been successfully used in the literature for anomaly detection in satellite imagery. For example, in [21], the authors employ a variational autoencoder architecture to detect anomalies in avalanche deposits in Synthetic Aperture Radar (SAR) satellite imagery. Similarly, the authors in [11] utilise the same methods to detect vessels in satellite imagery, identifying ships as anomalous data, thus dividing satellite imagery into smaller patches or tiles, which are then used for training the AE if they do not contain ships. It must be noted that this approach to vessel detection requires either: (I) satellite imagery without ships, or (II) segmentation labels to avoid patches containing ships when training the anomaly detection-based vessel detection AE.

Other approaches to anomaly detection in satellite imagery include generative adversarial networks (*GANs*). The original GAN architecture was formed by two different networks: (I) a generator, responsible for learning the data distribution, so that it can transform a noise vector z into a tensor that simulates the input data as closely as possible; and (II) a discriminator network responsible for detecting whether a given input comes from the real data distribution, or whether it was generated by generator network [22]. As can be understood, these two networks have opposing learning objectives, hence the technique's name. GANs were created for synthetic data generation, however, they can be used for anomaly detection and, in fact, many works employ GANs to carry out anomaly detection in many fields, including satellite imagery and EO [23].

The main inconvenience of the traditional approaches to GANs for anomaly detection is that they did not include an efficient manner to encode the anomaly detection target image into a noise vector that could then be fed to the generator for creating a synthetic input, allowing the discriminator to decide whether this input is an anomaly, based on the learned data distribution. More specifically, these traditional approaches were based on iteratively creating a vector from the original input, reconstructing such an input with the generator and using the reconstruction error with respect to the original input as the manner of improving and polishing the vector representation, which is significantly inefficient. To overcome this challenge, newer approaches include the encoding of the input data as part of the learning objective of the adversarial learning paradigm. An example of these improvements is the BiGAN, or Bi-directional Generative Adversarial Network, which includes an encoder as a third module of the architecture, so that encoder and generator can work together to fool the discriminator, and the discriminator tries to predict whether the input image comes from the true data distribution or whether it was generated by the encoder-generator modules [24].

Some other approaches to anomaly detection are partially derived from the aforementioned techniques. For example, the IZI and ZIZ architectures utilise a pre-trained GAN generator as the decoder of the AutoEncoder architecture for anomaly detection [13]. Both architectures are identical, except the order in which the operations are performed. On the one hand, the IZI architecture trains an encoder to learn the mapping of data from the input space i to the latent space z, and uses a pre-trained GAN generator to convert a vector in the latent space z to the input space i, utilising the reconstruction error as the usual anomaly score function. Conversely, the ZIZ architecture

utilises a pre-trained GAN generator to move from a noise vector z in the latent space to the input space i, and trains an encoder the latter into the latent space again z, so that the reconstruction error of the noise vector is used as the usual function to calculate the anomaly score [13], [25].

One of the most recent approaches to anomaly detection is the f-AnoGAN, which uses the existing IZI architecture and adds a pre-trained GAN discriminator to further refine the anomaly score function based on the features of the last layer[1] of the discriminator network [13]. This approach has been used in the context of anomaly detection in satellite imagery, achieving state-of-the-art results and improving the results of older approaches [13], [25], [26].

III. Methodology

This section describes the methods and the datasets used for the study, including the decision criteria for their selection. Furthermore, we also include a formal description of our proposed method, DualAnoDAE (see Section III-E).

A. Anomaly detection methods

In this work, we implemented different SSL anomaly detection methods, including some of the most recent approaches to anomaly detection in satellite imagery. More specifically, we implemented the AutoEncoder, IZI, ZIZ, BiGAN and f-AnoGAN anomaly detection methods, which use different architectures to solve the anomaly detection problem in the context of satellite imagery and EO.

The AutoEncoder was implemented according to the architecture specified in [13]. In their work, the authors use a simple convolutional AutoEncoder, characterized by three convolutional layers and RELU activations to encode the input image into a latent vector, which is used to carry out image reconstruction through three transposed convolutional layers and RELU activations. It must be noted that batch normalization layers are included to introduce regularization, reducing overfitting and improving the quality of the trained model. While alternative implementations of the AutoEncoder exist [21], [27], this implementation was selected due to its effective anomaly detection in satellite imagery [13].

Besides the AutoEncoder, the BiGAN anomaly detection architecture proposed in [24] was implemented according to the indications of the paper. The encoder and generator of the BiGAN were re-used from the AutoEncoder, and the discriminator was implemented according to [13]. This implementation is supported by state-of-the-art results achieved on anomaly detection in satellite imagery that the authors describe in their work using the same architecture [13].

Similarly, the IZI, ZIZ and f-AnoGAN architectures were also implemented according to their original papers. However, since all three models depend on a pre-trained GAN, the classical DCGAN proposed in [28] was implemented. The pre-trained generator of the implemented DCGAN was used as the decoder module of the all the three previously mentioned architectures. For the f-AnoGAN, the pre-trained discriminator of the DCGAN was also used, as specified by the authors in the original research paper.

Multiple reasons support the selection of these methods. First, all methods are self-supervised deep learning approaches to anomaly detection, which solves the curse of dimensionality described in Section II. Such a problem is not negligible in the EO domain because satellite acquisitions are usually extremely high-resolution imagery. On the other hand, the selected set of methods contains more classical approaches, like the AutoEncoder, as well as state-of-the-art methods, like the f-AnoGAN, thus providing an acceptable baseline for comparing and evaluating our method (see Section III-E).

B. Configuration

All of the aforementioned methods, including our novel method, were trained under the same settings. Using the same settings enables a fair comparison between methods that use similar – and in some cases the same – components, with minor details that differentiate each approach. Thus, using the same configurations allows us to establish a fair comparison and understand which approach provides better results for detecting anomalies in satellite imagery datasets. These training configurations, described in the following paragraph, are inspired by the work of the authors in [13], due to the similarity of the datasets used for their study compared to this work.

- Epochs: 200
- Learning rate: $1e-4$
- Patch size and format: 32x32 Planar RGB
- Batch size: 256
- Latent vector dimensionality: 1000

Although most of the configurations are inspired by the previous work, the batch size was set to 256 due to the memory constraints of the machine used to train the models when using patches of 32x32 pixels in RGB (see Section III-C). Although it is a controversial topic in the literature, some authors suggest that increasing the batch size provides a better estimation of the real error of the model, leading to better models [29]. In this case, the maximum batch size allowed in the host machine was 256, and hence the value of this setting was maximised to the limits of the available hardware.

Loss functions and anomaly score functions are also part of the configuration of the methods. For all the implemented methods, the loss functions and anomaly scores described in their original works was used. The summary of loss functions and anomaly scores is depicted in Table I, whereas the mathematical formulae are given in (1-4). The acronyms utilised in Table I are described next: BCE for Binary Cross Entropy, MSE for Mean Squared Error applied on the reconstructed input (either an image or a vector in the latent space), MAE for Mean Absolute Error applied on the reconstructed input, MSE_D for MSE applied on the features extracted at the last layer of the discriminator of the architecture for both the real image and the reconstructed image, and MAE_D for MAE applied on the aforementioned features of the discriminator.

[1]Any layer before the prediction, i.e. the sigmoid layer, can be used, although the last layer is recommended by the authors.

Furthermore, in (1-4), $\mathbf{x}$ refers to the matrix representation of the input image x. $\hat{\mathbf{x}}$ represents the model's reconstruction of the input image, $\mathbf{z}$ represents the latent vector, $\hat{\mathbf{z}}$ represents the model's reconstruction of the latent vector, $\mathbf{y}$ represents the ground truth labels used for training the GANs (i.e., whether the input image is a true image, or a fake image generated by the generator), and $\hat{\mathbf{y}}$ represents the model's prediction of the input image label. Last, $\mathbf{f}$ represents the discriminator features in the last layer obtained with the input image x, whereas $\hat{\mathbf{f}}$ represents the discriminator features in the last layer obtained with the reconstruction of the image x (i.e., $\hat{\mathbf{x}}$).

TABLE I
SOTA METHODS LOSS FUNCTIONS AND ANOMALY SCORES

Method	Loss function	Anomaly score
AutoEncoder	MSE	MSE
IZI	MSE	MSE
ZIZ	MSE	MSE
DCGAN	BCE	-
BiGAN	BCE	$MSE + MAE_D$
f-AnoGAN	$MSE + MSE_D$	$MSE + MSE_D$

$$MSE(\mathbf{x}) = \frac{1}{n}\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \tag{1}$$

$$MSE_D(\mathbf{x}) = \frac{1}{n}\|\mathbf{f} - \hat{\mathbf{f}}\|^2 \tag{2}$$

$$MAE(\mathbf{x}) = \frac{1}{n}\|\mathbf{x} - \hat{\mathbf{x}}\|_1. \tag{3}$$

$$MAE_D(\mathbf{x}) = \frac{1}{n}\|\mathbf{f} - \hat{\mathbf{f}}\|_1. \tag{4}$$

$$BCE(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i). \tag{5}$$

It must be noted that DCGAN and BiGAN use binary cross entropy losses, although the loss of generator (and encoder in the case of BiGAN) and discriminator modules are different and opposed, as expected from the adversarial learning paradigm. A full version of the mathematical definition of these losses is given in their original papers [24], [28]. Similarly, DCGAN does not have an associated anomaly score, because the DCGAN was trained to re-use its components for other architectures (e.g., IZI, ZIZ and f-AnoGAN), remaining unused for the purpose of anomaly detection due to the inefficient encoding of the input image into the latent space, described during Section II.

Image reconstruction error functions (e.g. MSE and MAE) have been widely used in the literature as loss and anomaly score functions, because they allow the model to learn the normal (i.e., non-anomalous) data distribution. Indeed, these model learn well how to reconstruct a normal image, having a harder time reconstructing anomalous imagery. This intuition, as well as state-of-the-art results, supports the usage of loss functions and anomaly scores grounded on the reconstruction error.

Last, the anomaly score threshold was computed as the average anomaly score of each model in the validation partition of the used datasets, as has been previously done in the literature, achieving good results [13], [30].

C. Datasets

Although numerous satellite imagery datasets exist, annotated datasets remain limited. A key advantage of self-supervised anomaly detection techniques over supervised methods (as discussed in Section II) is their ability to operate on non-annotated data. However, reliable evaluation of these methods requires annotated datasets, which enable the calculation of objective performance metrics such as accuracy and F1 score. For this study, we used LandCover.ai and HRC_WHU—two well-established, open-source satellite imagery datasets commonly used in research [31], [32].

Both datasets provide high-resolution satellite imagery with per-pixel annotations. LandCover.ai includes four labeled classes: buildings (1), woodland (2), water (3), and roads (4) [31]. HRC_WHU, on the other hand, consists of carefully annotated high-resolution imagery designed for developing cloud segmentation models using aerial sensor data [32]. Table II provides further details on the number of available satellite acquisitions and their dimensions.

TABLE II
DETAILS OF THE DATASETS USED FOR THE STUDY

Dataset	Satellite acquisitions	Width (px)	Height (px)
LandCover.ai	42	4200-9000	4700-9500
HRC_WHU	150	1280	720

Both datasets utilize high-resolution imagery, which is beneficial for many Earth Observation applications but presents challenges due to high memory consumption during inference. To address this, most studies divide satellite images into smaller patches or tiles [13]. The choice of tile size balances computational feasibility with model performance, as self-supervised anomaly detection methods rely on accurate image reconstruction [33]. Typically, patch sizes range from 28×28 to 64×64 pixels to optimize reconstruction quality and anomaly detection accuracy. In this study, a tile size of 32×32 pixels was selected to align with state-of-the-art methods and ensure comparability with existing approaches [13].

The datasets were partitioned into training (60%), validation (20%), and test (20%) sets. This allocation was determined based on the number of available patches for model training, ensuring sufficient data for learning and evaluation. Given the size of the datasets, techniques such as cross-validation considered unnecessary to train the models. Training and validation patches were processed through a filter to guarantee that only non-anomalous tiles were considered. In the case of LandCover.ai, water patches were considered the non-anomalous class, hence the model was trained with tiles were all pixels were annotated as water pixels. Similarly, in HRC_WHU patches with clouds were considered non-anomalous patches, whereas patches without clouds were considered an anomaly.

As a result of this filtering, the trained models were expected to learn the distribution of water patches for the case of the LandCover.ai dataset, and the distribution of cloud patches for the case of HRC_WHU.

As can be understood, the effectiveness of the implemented anomaly detection methods was evaluated based on their capability of detecting patches with different annotations than the ones used for training (e.g., road, woodland and building patches must be identified as anomaly patches by the anomaly detection models trained on LandCover.ai, whereas non-cloud patches must be identified as anomalous patches by the models trained on HRC_WHU). Further details on the evaluation process and the techniques used for assesing the effectiveness of the models can be found in Section III-D.

Figure 1 and Figure 2 present visual inspections of the datasets, showcasing 36 satellite acquisitions from the LandCover.ai dataset and 36 from the HRC_WHU dataset. Notably, even through visual assessment, LandCover.ai exhibits exceptionally high-resolution and high-quality imagery, as previously indicated in Table II, whereas HRC_WHU consists of satellite acquisitions with comparatively lower resolution. Given that these datasets serve distinct purposes—land cover prediction and cloud cover prediction, respectively—and differ in resolution, evaluating anomaly detection methods across both datasets can provide more comprehensive and robust conclusions.

D. Evaluation

When evaluating self-supervised anomaly detection methods on annotated datasets such as Landcover.ai and HRC_WHU, various metrics can be utilized. The evaluation process involves determining whether the trained anomaly detection models[2] correctly classify each patch as either anomalous or non-anomalous, based on the provided annotations. Consequently, classification metrics such as accuracy, precision, and recall play a fundamental role in assessing the performance of anomaly detection models, and they have been used in many works on satellite imagery anomaly detection [13], [34].

For this work, we selected accuracy, precision, recall, and F1 score as the metrics with which to evaluate the implemented models, including the state-of-the-art models and DualAnoDAE. This selection seems to provide different perspectives of the behaviour of the model when running inference on the test partition of each dataset. Accuracy provides the percentage of patches that were correctly classified, however, since the datasets are not balanced, a model predicting the most frequent class would always be perceived as a better model by this metric. For these reasons, precision, recall, and F1 were also included. On the one hand, precision evaluates whether the models are reliable in their predictions (i.e., if they make an anomaly prediction, the patch will likely be an anomaly), whereas recall evaluates whether the model is capable of identifying all anomalous patches, regardless of the number of anomalous patch predictions. Since precision and recall offer useful and different perspectives, the F1 score metric is also used for the evaluation, as it combines precision and recall to provide a score that is useful to evaluate models, even if they were trained on imbalanced datasets.

[2]It is important to note that these models were not exposed to anomalous patches during training.

E. DualAnoDAE

In this study, we conducted a comprehensive evaluation of state-of-the-art algorithms for anomaly detection in satellite imagery. While most methods demonstrated strong performance in identifying anomalies (see Section IV), a critical limitation emerged: all approaches struggled with false negatives, ultimately reducing their recall. These false negatives—where genuinely anomalous patches were misclassified as non-anomalous—highlight a fundamental issue. Specifically, the ability of these methods to reconstruct previously unseen images results in anomaly scores that fall below the detection threshold, causing the system to overlook anomalies. This finding underscores a key challenge in current anomaly detection techniques and suggests the need for more robust strategies to mitigate false negatives.

In this work, we propose a Dual Anomaly Detection AutoEncoder (*DualAnoDAE*), an ensemble of two autoencoders aimed at learning both general and nuanced features of the input image to help its reconstruction. The intuition behind DualAnoDAE is to use a two-fold approach to anomaly detection in satellite imagery: (I) a high-level of abstraction AutoEncoder is included with a high receptive field, to learn general features of the input, non-anomalous, patches. Conversely, (II) a low-level of abstraction AutoEncoder is included with a low receptive field, to learn very nuanced features of the non-anomalous tiles. By using an ensemble of two AutoEncoders with different receptive fields, DualAnoDAE possesses more information about the input patch than state-of-the-art methods, ultimately improving its effectiveness, reducing the number of false positives and negatives, and surpassing the results of state-of-the-art methods.

The high-level AutoEncoder is designed with three convolutional blocks, each utilizing a 5×5 px receptive field and pooling layers. This architecture, inspired by the AutoEncoder proposed in [13] but with modified receptive fields, ensures that the final convolutional layer effectively captures nearly the entire 32×32 patch during training. As a result, this AutoEncoder focuses on extracting broad structural features to reconstruct the input image. In contrast, the low-level AutoEncoder employs convolutional networks with a 3×3 px receptive field, enabling it to capture finer-grained details with greater precision. By combining both models, the system leverages high- and low-level abstraction features, leading to a more robust and accurate anomaly detection approach for satellite imagery, where capturing both global structures and intricate patterns is essential for identifying anomalies.

A description of the architecture of DualAnoDAE is shown in Fig. 3, where the upper AutoEncoder represents the high-level of abstraction autoencoder, that uses its high receptive

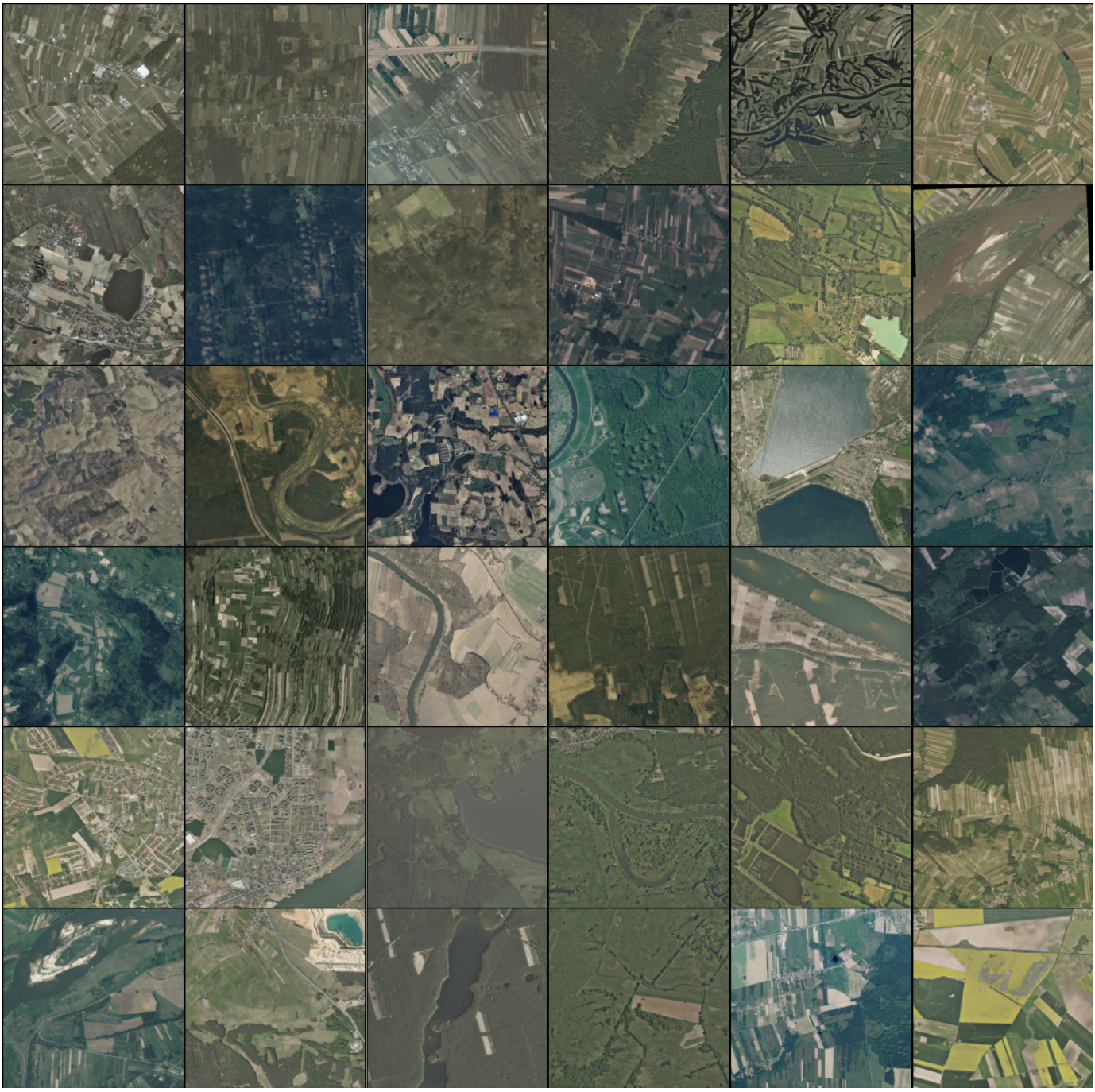

Fig. 1. Visual inspection of the LandCover.ai dataset

field to reconstruct the input patch (X), generating a high-level reconstruction of the input ($\hat{X_{high}}$. Similarly, the AutoEncoder in the lower part of the image represents the low-level of abstraction AutoEncoder, which uses its low receptive field to reconstruct the input image into a low-level of abstraction reconstruction ($\hat{X_{low}}$).

Our novel method generates two input reconstructions for a given input patch. The two reconstructions are used for anomaly detection, as the anomaly score is defined by the formula given in (6), where X represents the input patch, and $\hat{X_{high}}$ and $\hat{X_{low}}$ represent the input reconstructions of the high- and low-level of abstraction AutoEncoders respectively.

$$\mathcal{A}(X) = \frac{1}{n}\|X - \hat{X}_{low}\|^2 + \frac{1}{n}\|X - \hat{X}_{high}\|^2 \qquad (6)$$

The anomaly score threshold was defined in the same way as the state-of-the-art-methods: we used the average anomaly score value obtained from the patches in the validation dataset

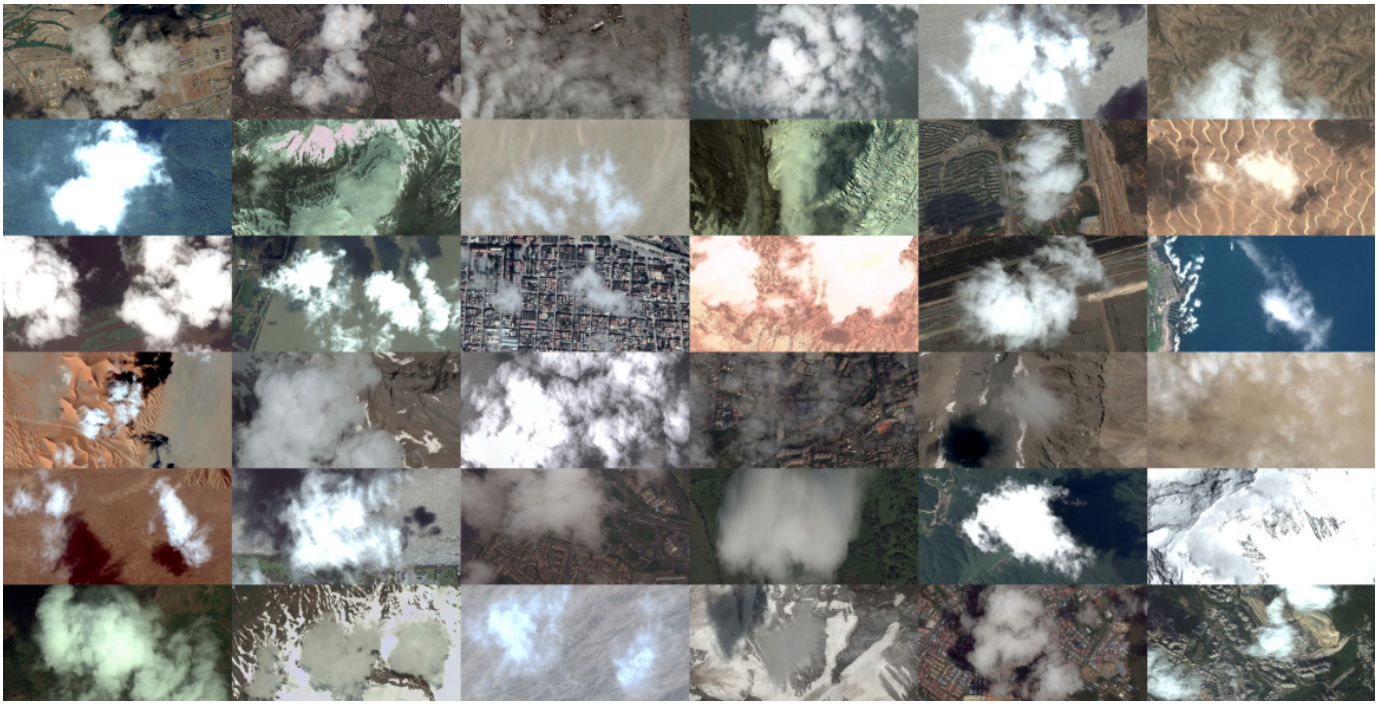

Fig. 2. Visual inspection of the HRC_WHU dataset

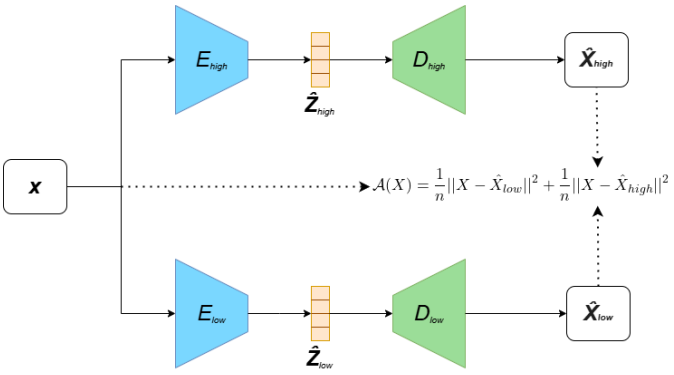

Fig. 3. DualAnoDAE architecture

as the anomaly score threshold to be used during the evaluation of the method.

Last, both AutoEncoders were trained separately, using the loss function defined in Table I, thus using settings that are as similar as possible to the ones used when training and evaluating the state-of-the-art methods, providing a reliable comparison between them.

IV. Results

As explained in Section III, all the state-of-the-art anomaly detection methods, as well as our novel method, DualAnoDAE, were trained on the two selected satellite imagery datasets: LandCover.ai and HRC_WHU. Table III contains the results of each implemented anomaly detection method for the selected evaluation metrics on the LandCover.ai dataset, whereas Table IV contains the results obtained by all models in the HRC_WHU dataset.

TABLE III
EVALUATION METRICS PER METHOD IN THE LANDCOVER.AI DATASET

Model	Accuracy	Precision	Recall	F1
AutoEncoder	0.7891	0.7685	**0.9738**	0.8590
IZI	0.9267	0.9730	0.9414	0.9569
ZIZ	0.7445	0.9714	0.7260	0.8310
BiGAN	0.7888	0.9602	0.7886	0.8660
f-AnoGAN	0.9301	0.9745	0.9438	0.9589
DualAnoDAE	**0.9477**	**0.9838**	0.9552	**0.9693**

TABLE IV
EVALUATION METRICS PER METHOD IN THE HRC_WHU DATASET

Model	Accuracy	Precision	Recall	F1
AutoEncoder	0.7788	0.7633	0.9637	0.8518
IZI	0.7753	0.7610	0.9615	0.8496
ZIZ	0.7349	0.7910	0.8131	0.8019
BiGAN	0.7175	**0.8023**	0.7589	0.7800
f-AnoGAN	0.7763	0.7615	0.9624	0.8503
DualAnoDAE	**0.7958**	0.7757	**0.9714**	**0.8626**

The results presented in Table III indicate that, in the LandCover.ai dataset, DualAnoDAE achieved the highest Accuracy (0.9477), Precision (0.9838), and F1-score (0.9693), and a second-best Recall (0.9552). These results suggest that the model is not only capable of correctly identifying anomalies but also maintaining a low false positive rate (high precision). While f-AnoGAN and IZI also performed well, achieving F1-scores of 0.9589 and 0.9569, respectively, DualAnoDAE exhibited a slight yet meaningful improvement, reinforcing its robustness in this dataset.

The HRC_WHU dataset (Table IV) poses more challenges, due to cloud patches being considered non-anomalous patches. The previous decision leads to a harder anomaly detection problem because not all clouds remain fully opaque, thus the models can learn how to reconstruct tiles with nearly-transparent clouds, which in turn can help models reconstruct non-cloud (i.e., anomalous) patches. Regarding the results, DualAnoDAE achieved the highest Accuracy (0.7958) and F1-score (0.8626), outperforming traditional AutoEncoders, GAN-based methods, and IZI. Notably, its Recall (0.9714) was the highest among all models, suggesting that DualAnoDAE is particularly effective at identifying true anomalies in this dataset, showcasing the effectiveness of the intuition behind our proposed method.

Considering the results, f-AnoGAN and DualAnoDAE demonstrate the highest accuracy in anomaly detection, significantly improving recall—a key metric closely linked to the reduction of false negatives. Notably, their superior performance is not solely attributed to their architectural design but rather to the extent of information they leverage. This is evident in f-AnoGAN, which integrates knowledge from multiple sources, including the discriminator, a pre-trained decoder, and its own trained encoder, maximizing the information available for anomaly detection. Similarly, DualAnoDAE benefits from an ensemble of two AutoEncoders with significantly different receptive fields, enabling it to use general and nuanced features that boost its performance. These findings suggest that incorporating more information, rather than relying on a specific type of model architecture, is the key factor in achieving state-of-the-art results in the field of anomaly detection in satellite imagery.

Finally, Fig. 4 presents a visual assessment of the anomaly detection results. The figure displays the outputs of different anomaly detection methods applied to four randomly selected images from the test partition of both datasets. More specifically, the first four satellite acquisitions belong to the LandCover.ai dataset, whereas the last four satellite acquisitions belong to the HRC_WHU dataset. For the former dataset, water was considered the non-anomalous class (black color), whereas any non-water patch should be considered an anomalous patch (white color).

The visual inspection of the results further supports the robustness of f-AnoGAN and DualAnoDAE, as indicated by the obtained metrics. These methods consistently outperform simpler approaches in anomaly detection, as evident in their visual outputs. Specifically, we observe that: (I) AutoEncoder and IZI exhibit a significant number of false negative predictions across both datasets, a limitation that is particularly pronounced in the LandCover.ai dataset, where these models misinterpret forest patches as greenish water, leading to incorrect non-anomalous classifications. (II) Similarly, ZIZ and BiGAN also struggle with false negative detections, but to an even greater extent. This can be attributed to their ability to generate reconstructed images that closely resemble the original data distribution, a characteristic influenced by their training process using noisy, i.e., random, vectors. (III) Finally, f-AnoGAN and DualAnoDAE demonstrate superior anomaly detection performance across all satellite imagery. Although certain images remain challenging—such as forest patches resembling greenish water—these methods exhibit greater robustness, substantially reducing false negative detections and improving overall result quality. Undoubtedly, their robustness can be attributed to the usage of additional information when running anomaly detection tasks, which is particularly evident in DualAnoDAE, a novel method that uses an ensemble of two AutoEncoders with very different receptive fields to gather both general and nuanced features, with which to simplify the anomaly detection task, even if it is performed in complex and high-resolution satellite imagery.

V. Limitations and future work

This study acknowledges several limitations that open the door to future research. One of the primary challenges is the vast and continuously evolving landscape of state-of-the-art anomaly detection techniques in satellite imagery. While a set of five different methods was implemented for comparison, other approaches exist, each with unique strengths and weaknesses. Future work could explore additional techniques to further validate the effectiveness of the proposed DualAnoDAE model. These novel approaches include transformer models, which seem to be fairly capable at detecting anomalies in other fields, yet they pose challenges, particularly in terms of computational efficiency and deployment feasibility on resource-constrained devices, like satellites, which have strict limitations on power consumption, memory, and processing capacity, making it difficult to integrate such high-complexity models without significant optimization efforts.

Furthermore, the study relied on two datasets for evaluating the proposed approach. While these datasets provided valuable insights, anomaly detection performance may vary under different circumstances. Expanding the evaluation to include a broader range of datasets could enhance the generalizability and robustness of the findings, ensuring that our novel method performs well across diverse real-world scenarios.

Additionally, we observed that certain satellite acquisitions present challenges for self-supervised anomaly detection methods. Specifically, under particular circumstances—such as forest patches resembling greenish water, even to the human eye—these methods can reconstruct anomalous patches with a reconstruction error below the average threshold, resulting in an increased number of false negatives. While this limitation is inherent to the nature of self-supervised approaches, we believe that future research focused on enhancing the information available during the anomaly detection process could significantly improve performance. This potential for improvement is exemplified by our novel method, DualAnoDAE.

Lastly, we recognize the potential of self-supervised anomaly detection techniques in addressing related challenges within the broader field of Earth Observation. In particular, future research could explore the effectiveness of these methods in autolabeling, a machine learning approach where

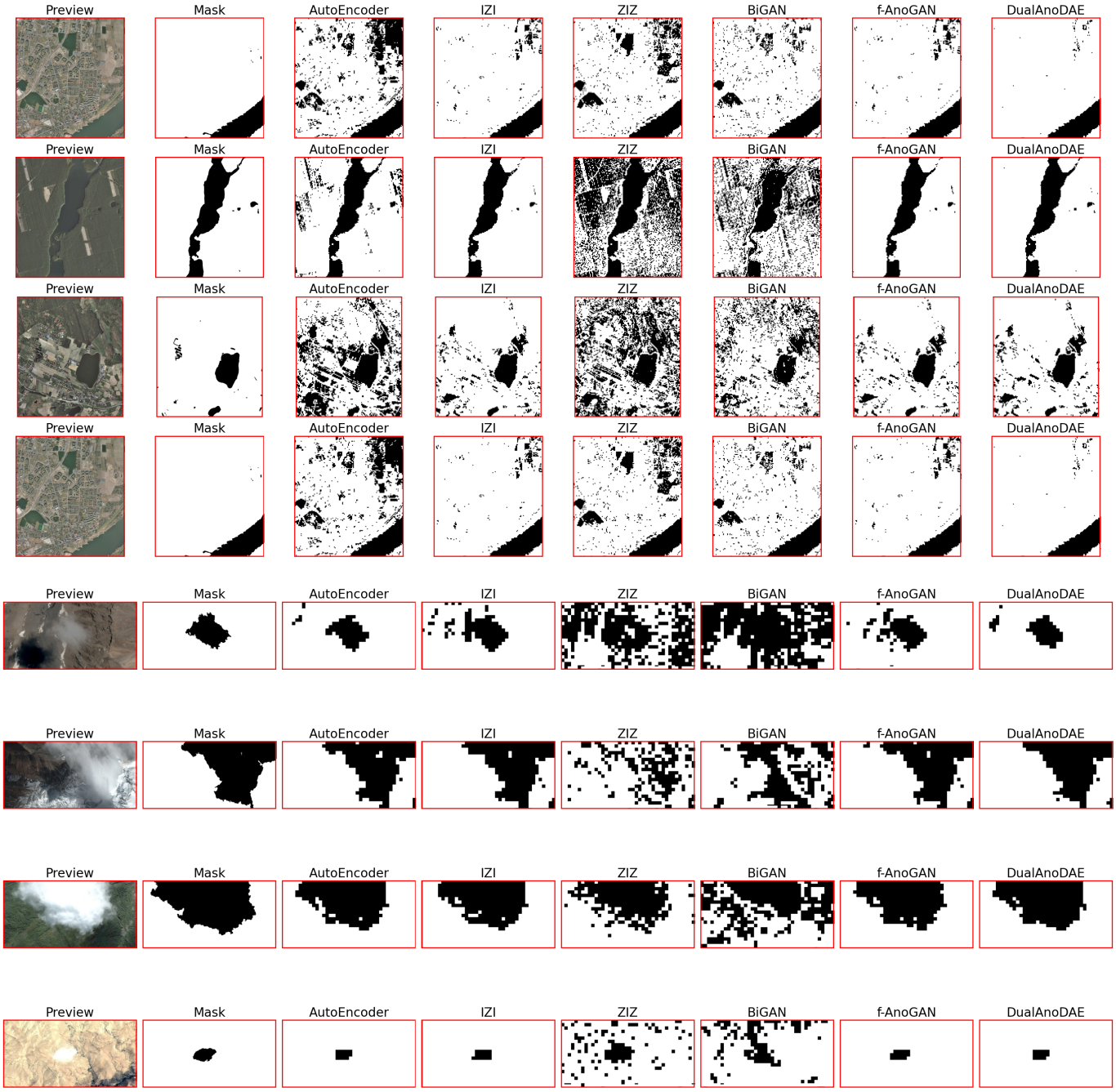

Fig. 4. Visual inspection

models learn to segment input images after training on non-annotated data. This could significantly reduce the dependency on manually labeled datasets, which are often scarce and expensive to produce.

VI. Conclusions

In this work, we explored self-supervised anomaly detection methods for satellite imagery and evaluated their effectiveness on two well-established datasets: LandCover.ai and HRC WHU. Traditional supervised approaches to anomaly detection often face limitations due to the scarcity and high cost of labeled datasets. By leveraging self-supervised learning, multiple works in the literature demonstrated that effective anomaly detection can be performed without requiring explicit anomaly labels, making it a feasible solution for real-world applications in the field of EO.

Our study compared multiple state-of-the-art methods, including AutoEncoder, IZI, ZIZ, BiGAN, and f-AnoGAN,

providing a comprehensive assessment of their behaviour using accuracy, precision, recall, and F1-score as evaluation metrics. While these methods showed competitive results, we identified key weaknesses, particularly in their ability to reduce the number of false positives and, mainly, false negatives.

To address these limitations, we proposed a novel self-supervised anomaly detection method, DualAnoDAE. Our method incorporates both high- and low-level feature extraction through a dual-autoencoder architecture, enabling it to capture both general and nuanced features, which can be used to detect anomalies with improved robustness. The experimental results demonstrated that DualAnoDAE surpasses existing techniques for two different satellite imagery datasets, and two different learning objectives (non-water anomaly detection and non-cloud anomaly detection).

Despite these promising results, our study acknowledges certain limitations. The evaluation was conducted on two datasets, and while they offer diverse geospatial contexts, additional datasets should be explored to further validate the generalizability of our approach. Additionally, future research could investigate more advanced techniques, such as transformer-based anomaly detection models, to further enhance performance. Lastly, real-time applicability and deployment on resource-constrained platforms, such as satellite onboard processing systems, remain open areas of research within the field of anomaly detection in satellite imagery.

Overall, this study contributes to anomaly detection in satellite imagery by benchmarking existing state-of-the-art approaches on two different satellite imagery datasets, identifying some of their limitations, and introducing a novel and effective self-supervised method that provides promising results to address them. Future research should focus on further enhancing self-supervised learning for Earth observation, optimizing these approaches for deployment on resource-constrained platforms such as satellites, and exploring their potential for autolabeling to address the limited availability of high-quality annotated datasets for anomaly detection.

ACKNOWLEDGMENT

I want to thank both Jorge Díez Peláez and José Luis Espinosa Aranda for their invaluable help in developing this work. I have had the luck to learn from two great experts in the topic who have not only supported me during the Master's Thesis, but also greatly contributed to increasing my passion towards the development of AI systems that can be helpful for society.

REFERENCES

[1] X. Yang, J. Blower, L. Bastin, V. Lush, A. Zabala, J. Masó, D. Cornford, P. Díaz, and J. Lumsden, "An integrated view of data quality in earth observation," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1983, p. 20120072, 2013.

[2] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon, "Using satellite imagery to understand and promote sustainable development," *Science*, vol. 371, no. 6535, p. eabe8628, 2021.

[3] S. M. Pekkanen, S. Aoki, and J. Mittleman, "Small satellites, big data: uncovering the invisible in maritime security," *International Security*, vol. 47, no. 2, pp. 177–216, 2022.

[4] Z. Hong, Z. Tang, H. Pan, Y. Zhang, Z. Zheng, R. Zhou, Z. Ma, Y. Zhang, Y. Han, J. Wang *et al.*, "Active fire detection using a novel convolutional neural network based on himawari-8 satellite images," *Frontiers in Environmental Science*, vol. 10, p. 794028, 2022.

[5] J. Mendoza-Bernal, A. Gonzalez-Vidal, and A. F. Skarmeta, "A convolutional neural network approach for image-based anomaly detection in smart agriculture," *Expert Systems with Applications*, vol. 247, p. 123210, 2024.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[7] J. Imbert, G. Dashyan, A. Goupilleau, T. Ceillier, and M.-C. Corbineau, "Improving performance of aircraft detection in satellite imagery while limiting the labelling effort: Hybrid active learning," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 220–223.

[8] H. Zhao, M. Liu, S. Qiu, and X. Cao, "Satellite unsupervised anomaly detection based on deconvolution-reconstructed temporal convolutional autoencoder," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2989–2998, 2023.

[9] M. J. Hashemi, E. Keller, and S. Tizpaz-Niari, "Detecting unseen anomalies in network systems by leveraging neural networks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 2515–2528, 2022.

[10] Z. Lin, H. Wang, and S. Li, "Pavement anomaly detection based on transformer and self-supervised learning," *Automation in Construction*, vol. 143, p. 104544, 2022.

[11] A. J. Farr, I. Petrunin, G. Kakareko, and J. Cappaert, "Self-supervised vessel detection from low resolution satellite imagery," in *AIAA SCITECH 2022 Forum*, 2022, p. 2110.

[12] M. Munir, M. A. Chattha, A. Dengel, and S. Ahmed, "A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2019, pp. 561–566.

[13] M. A. Contreras-Cruz, F. E. Correa-Tome, R. Lopez-Padilla, and J.-P. Ramirez-Paredes, "Generative adversarial networks for anomaly detection in aerial images," *Computers and Electrical Engineering*, vol. 106, p. 108470, 2023.

[14] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[17] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Irish conference on artificial intelligence and cognitive science*. Springer, 2009, pp. 188–197.

[18] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9664–9674.

[19] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer, "Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels," *Medical Image Analysis*, vol. 78, p. 102385, 2022.

[20] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[21] S. Sinha, S. Giffard-Roisin, F. Karbou, M. Deschatres, A. Karas, N. Eckert, C. Coléou, and C. Monteleoni, "Variational autoencoder anomaly-detection of avalanche deposits in satellite sar imagery," in *Proceedings of the 10th International Conference on Climate Informatics*, 2020, pp. 113–119.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[23] M. Sabuhi, M. Zhou, C.-P. Bezemer, and P. Musilek, "Applications of generative adversarial networks in anomaly detection: A systematic literature review," *Ieee Access*, vol. 9, pp. 161 003–161 029, 2021.

[24] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[25] C. Park, S. Lim, D. Cha, and J. Jeong, "Fv-ad: F-anogan based anomaly detection in chromate process for smart manufacturing," *Applied Sciences*, vol. 12, no. 15, p. 7549, 2022.

[26] O. Sliti, M. Devanne, S. Kohler, N. Samet, J. Weber, and C. Cudel, "f-anogan for non-destructive testing in industrial anomaly detection," in *Sixteenth International Conference on Quality Control by Artificial Vision*, vol. 12749. SPIE, 2023, pp. 297–304.

[27] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 665–674.

[28] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[29] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry, "Augment your batch: better training with larger batches," *arXiv preprint arXiv:1901.09335*, 2019.

[30] A. Komadina, M. Martinić, S. Groš, and Ž. Mihajlović, "Comparing threshold selection methods for network anomaly detection," *IEEE access*, 2024.

[31] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 1102–1110.

[32] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 197–212, 2019.

[33] J. Horváth, D. Güera, S. K. Yarlagadda, P. Bestagini, F. M. Zhu, S. Tubaro, and E. J. Delp, "Anomaly-based manipulation detection in satellite images," *networks*, vol. 29, no. 21, pp. 62–71, 2019.

[34] A. Gulenko, M. Wallschläger, F. Schmidt, O. Kao, and F. Liu, "Evaluating machine learning algorithms for anomaly detection in clouds," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 2716–2721.