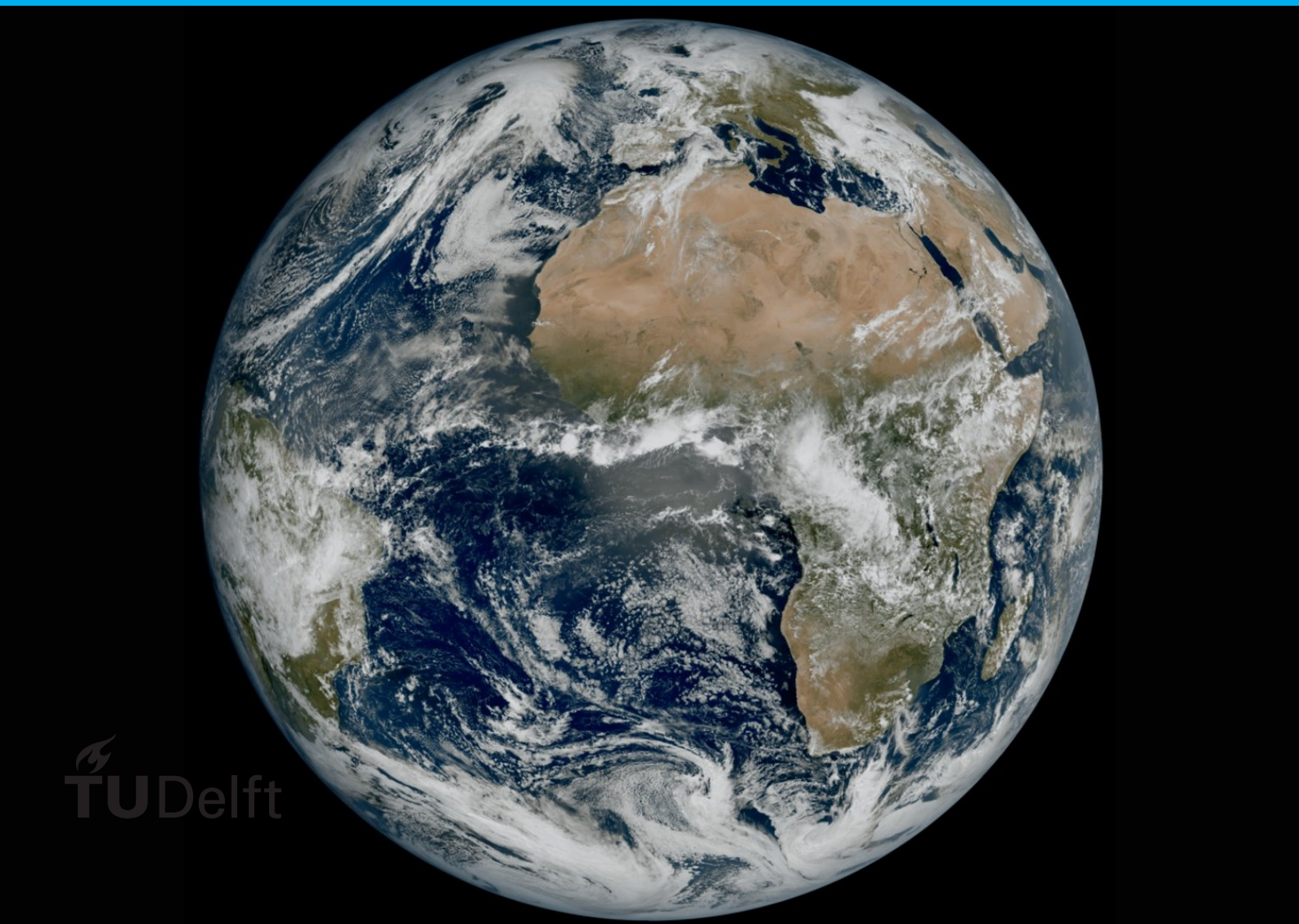
Anomaly Detection in Geostationary Satellites

Unsupervised, Satellite-Agnostic Error Detection & Localisation

Marco Bak

Master of Science Thesis
Computer Science, EEMCS
Conducted in collaboration with S[&]T

June 2026



TU Delft

Anomaly Detection in Geostationary Satellites

Unsupervised, Satellite-Agnostic Error Detection & Localisation

by

Marco Bak

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday, June 15, 2026 at 10:00 AM.

Student number:	5066212	
Project duration:	November 10, 2025 - June 15, 2026	
Thesis committee:	Dr. N. Yorke-Smith,	TU Delft, supervisor
	Dr.ir. J.G. Teixeira Encarnação,	TU Delft
	Sytze Andringa,	S[&]T
	Erwin Platen,	S[&]T

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Acknowledgements

This thesis was written in collaboration with S[&]T and the Delft University of Technology. I would like to thank my supervisors at S[&]T, Sytze Andringa and Erwin Platen, for their knowledge, guidance, and support throughout our weekly meetings. At the Delft University of Technology, I am grateful to Dr. Ir. Neil Yorke-Smith for his valuable feedback and insights, and Dr.ir. J.G. Teixeira Encarnação, for his willingness to be part of the thesis committee. I also want to thank my colleagues at S[&]T for making me feel welcome and turning my time at the company into an enjoyable and educational experience. Finally, I thank my family and friends for their encouragement and support during the writing of this thesis.

Abstract

This thesis studies the application of deep learning techniques for anomaly detection in geostationary satellite imagery. The focus is on developing an unsupervised, satellite-agnostic approach that can effectively identify and localise anomalies without relying on labelled data. The method uses f-AnoGAN to learn the underlying distribution of nominal satellite images and detect anomalies. It is designed to work across different satellites, spectral bands, and acquisition conditions. Improvements to the base model are investigated, including tile overlap for more global context and metadata conditioning. The generalisation of the model to unseen satellites via transfer learning is also evaluated. The results show that the model is able to detect a range of anomalies in the satellite images, with varying performance depending on the anomaly type. Fine-grained pixel anomalies and misalignment anomalies are more difficult to detect than structural anomalies like broken scan lines and noise patterns. While the base model achieved an AUROC of 0.7756, the overlap and metadata-augmented models showed improvement in detection performance, reaching 0.8155 and 0.8071 respectively. The transfer-learning experiment showed that training a new model from scratch on the target domain outperforms fine-tuning a pre-trained model. These results show that the f-AnoGAN method only partially meets the requirements of a standalone anomaly detection method. However, it shows promise as a first-pass method to flag novel anomalies in unseen datasets, enabling faster recognition of new anomalies and the creation of new rule-based detectors.

Fair use of AI

During the process of writing this thesis, AI tools were used in a safe and ethical way. The tools were used to assist the author in writing and implementing the methods. However, it is important to know that all the ideas and concepts described in this thesis are the result of the author's own effort and creativity, building on the foundations laid out in the background and related work sections. Used AI tools:

- Claude code for implementing evaluation methods, visualisation, data processing.
- Claude for generating tables from results, formulas from the code and for generating the outline of the methods' architecture in the text
- Claude for refining and reviewing the written text, and the flow of the thesis document.

By providing open information about the use of AI tools, the author wants to show that when used responsibly, AI is a strong tool to reinforce research, but not take over the goal of learning and doing research.

Contents

1 Introduction

1.1. Problem Definition

Over the last four decades, much research and engineering effort has been put into earth observation, with one of the main goals being understanding and predicting weather and climate. For this, satellites making full earth observations on multiple wavelengths are playing an important role. To ensure the quality, reliability and accuracy of this research, it is important to have access to reliable, high quality data. Unfortunately, satellite images are not error free. Errors are caused by various reasons, such as failing hardware, sunlight interfering with instruments or data transmission problems are causing errors in large parts or only a few pixels of the images.

With the growth of data-driven climate research, fast and reliable data processing methods become increasingly important to support this demand. This means that satellite images, often used in long-term climate and weather research, are crucial to meet this demand. Because of the long time span that data has already been generated, ensuring good data quality is important as it provides more confidence in research results.

In this thesis, the focus will be on data from geostationary satellites. Unlike low-earth orbit satellites, geostationary satellites continuously image the same region at short intervals over decades, producing large, long-running archives that are particularly difficult to validate manually and therefore well-suited to automated anomaly detection.

Different space agencies have launched multiple generations of geostationary satellites, each with different instruments, and slightly different spectral bands. Due to differences in instruments across and within satellite programmes, anomaly detectors must be tuned for each data format. Newer satellites often carry more spectral bands or upgraded instruments, which can introduce previously unseen anomaly types. Over time many anomalies have already been identified in the different datasets. For all of these anomalies, different detectors and repair algorithms were created. This required experts to look into the data and satellite instruments to actively create new, or tune existing anomaly detectors for known and new anomalies, making the process of anomaly detection and repair a time-consuming process.

Having a general detector capable of detecting a wide range of anomalies in data from multiple different generations of satellites would streamline the preparation of newly acquired satellite imagery. Such an algorithm would be able to identify most of the present anomalies. This means that only after the anomalies have been detected and localised an expert would have to confirm the found anomalies, and verify the found anomalies that have low confidence.

The goal of this research is to investigate and design **generic anomaly detection methods**. In the best case, this method would be able to detect a wide range of anomalies, varying from pixel to image level. Furthermore, it should be able to recognise anomalies in images of different types of satellites. The different satellites capture images on different spectral bands, where the satellites also capture images in different resolutions.

The current process for anomaly detection is as follows:

1. Random images are sampled from each satellite in the new dataset.
2. The sampled images are manually checked by experts for anomalies.

3. All found anomalies are classified.
4. For each anomaly type, a specific rule-based algorithm is designed and implemented.
5. The complete dataset is checked using these algorithms.

This process can take up quite some time, due to the manual work required. Furthermore, checking only a subset of the data can be error-prone, in particular for rare anomalies. Another issue is that for the newer datasets, with images up to 22,000 × 22,000 pixels, it becomes increasingly hard to spot line- or pixel-level anomalies. This process can be improved by creating a general anomaly detector that can:

- Detect rare novel anomalies in new datasets;
- Detect if an image contains any kind of anomaly, not just one;
- Create a mask covering the detected anomalies; and
- Work across different kinds of satellites without satellite-specific tuning.

From these requirements, the research questions are formulated, and discussed in the next section.

1.2. Research Question

Primary Research Question:
How can anomaly detection methods be designed to detect both pixel- and image-level anomalies in different geostationary satellite imaging systems without requiring labelled anomaly examples?

Note that satellite metadata (spectral band, acquisition time, tile position) is treated as part of the input signal rather than as satellite-specific supervision, since it is automatically available for every image without human annotation. From this primary research question, the following sub-questions are derived:

Sub-questions:

1. What anomaly detection algorithms are able to detect anomalies of different nature (smaller anomalies like hot pixels and bigger anomalies like corrupt lines) in geostationary satellite images?
2. How can anomaly detection across generations of GOES, Himawari and Meteosat satellites be handled?
3. Where does a learned anomaly detector complement the rule-based GIAD approach, and how can the two methods support each other in an anomaly detection pipeline?

1.3. Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 provides background on the satellite datasets and the known anomaly types present in them, as well as the current rule-based detection methods. Chapter 3 reviews related work on anomaly detection methods and their applicability to this problem. Chapter 4 describes the proposed methods. Chapter 5 presents the experimental setup and results. Chapter 6 discusses the results and their implications. Chapter 7 concludes with answers to the research questions and directions for future work.

2

Background

In this chapter, the context and background in which this research is conducted are explained. It starts by giving a short introduction to remote sensing with geostationary satellites, including the dataset used in this research. Next, the different anomalies found in the dataset are discussed, and nominal images are defined. Lastly, the current rule-based methods used for anomaly detection in satellite imagery are discussed.

2.1. Remote sensing and geostationary satellites

Geostationary satellites orbit at around 36,000 km from the Earth's equator, and their movement speed matches the Earth's rotation, keeping them fixed over a specific point on the Earth's surface. This provides the advantage for remote sensing that the satellite always views the Earth from the same perspective, enabling it to record the same region at brief intervals. This arrangement is particularly useful for observations of weather conditions [1]. Figure 2.1 shows the locations of the different geostationary weather satellites. The GOES satellites are positioned over the Americas, the Meteosat satellites are positioned over Europe and Africa, and the Himawari satellites are positioned over Asia and Australia. The dataset available for this thesis contains

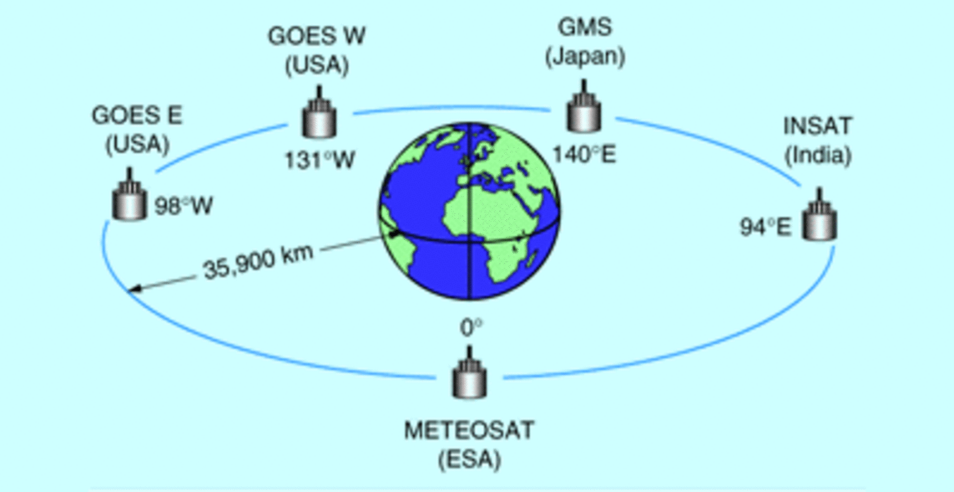


Figure 2.1: Locations of the different geostationary weather satellites [1].

data from three geostationary weather satellite programmes operated by different space agencies. Each programme has evolved over multiple generations, with each generation introducing improved imaging instruments that capture a different number of spectral bands at slightly different wavelengths [2, 33, 34]. Table 2.1 gives an overview of the programmes and their generations.

These satellites are equipped with various instruments such as SEVIRI in Meteosat, ABI in GOES, and AHI in Himawari [50, 32]. The instruments capture images in multiple spectral bands, including visible light and infrared. A spectral band is a specific range of wavelengths in the electromagnetic spectrum. This means

Table 2.1: Overview of satellite programmes and generations in the dataset.

Programme	Agency	Generation	Satellites	First launch
Meteosat	EUMETSAT	MFG	Meteosat 1–7	1977
		MSG	Meteosat 8–11	2002
		MTG	Meteosat 12+	2022
GOES	NOAA	Pre-GVAR	SMS 1–2, GOES 1–7	1975
		GVAR	GOES 8–15	1994
		GOES-R	GOES 16–19	2016
Himawari	JMA	GMS	Himawari 1–5	1977
		MTSAT	Himawari 6–7	1999
		AHI	Himawari 8–9	2014

that images taken across different spectral bands can look very different; visible light and near infrared bands are affected by day/night cycles, while water vapour and thermal infrared bands are not, some bands show highly detailed cloud structures, while others show more smoothed patterns [13]. Besides the visual differences, these different spectral bands allow the satellites to capture various features of the Earth's atmosphere and surface, such as cloud cover, temperature, moisture content, vegetation, ice cover, and more. The data from these satellites is crucial for weather forecasting, climate monitoring, and various scientific research applications [55, 57]. In this thesis, the channels will be referred to as spectral bands, more specifically as infrared (IR) which are also split into near- and thermal infrared (NIR and TIR), visible (VIS), and water vapour (WV) bands, which are commonly used in weather satellites. The water vapour bands are a specific type of infrared band that is sensitive to the presence of water vapour in the atmosphere, making it particularly useful for monitoring moisture and cloud formation.

Every image in the dataset can be associated with a specific spectral band (channel) and satellite. This information can be used as metadata to help a detector learn to associate certain image features with specific spectral bands or satellites.

2.2. Nominal images

In this thesis, a nominal image is defined as an image that does not contain anomalies that can be detected by the rule-based algorithms in GIAD (Section 2.4). An exception to this definition is the misalignment of the earth disk, where the earth disk is shifted in the image. This is a common anomaly, but does not show defects in the image. Due to the high frequency and the low impact, this anomaly is not considered for this thesis, and is therefore not included in the definition of an anomaly. Figures 2.2 and 2.3 show examples of nominal images from the Meteosat-9 and MTSAT-1R satellites, across three spectral bands. These images are free from detectable anomalies and represent the expected appearance of satellite imagery under normal conditions. When looking at the nominal images in Figures 2.2 and 2.3, it shows that there are small differences between the different satellites. The main difference is the part of the earth that is captured by the satellite. The MET9 captures the African landmass, where the MTSAT1 captures the Australian landmass. There are also some differences in how the background of the image is processed.

2.3. Known anomalies in the dataset

To be able to identify and detect as many anomalies as possible, it is important to understand the different anomalies that can occur in the dataset. There are geometric errors, where there is an error in the geometry of the earth. Here scanlines have the correct values, but are shifted so that the true geometric location does not overlap, as displayed in Figure 2.4a. Geometric errors can also be that the whole earth disk is shifted in the image. The other kind of errors are radiometric errors, more in depth described in Poulet [42] and Stassinopoulos et al. [56], where the error occurs because of errors occurring in the instrument and where the pixel values are incorrect. An example is shown in Figure 2.4b. Here we see a few lines that have a higher brightness than normal, and also a line with random noise pixels. Another interesting radiometric anomaly is anomalies caused by straylight. These anomalies are caused by sunlight falling on the imaging instruments, causing brighter spots on the image [42]. An example is shown in Figure 2.4c. Straylight anomalies are most likely to happen around equinoxes, because then the sun is most aligned with the satellite, giving a higher

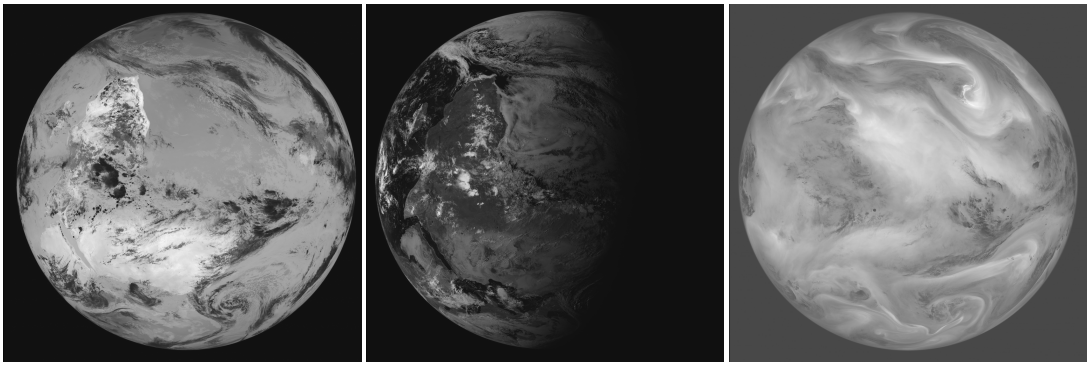


Figure 2.2: Nominal Meteosat-9 images across three spectral bands. From left to right: infrared (IR), visible (VIS), and water vapour (WV).

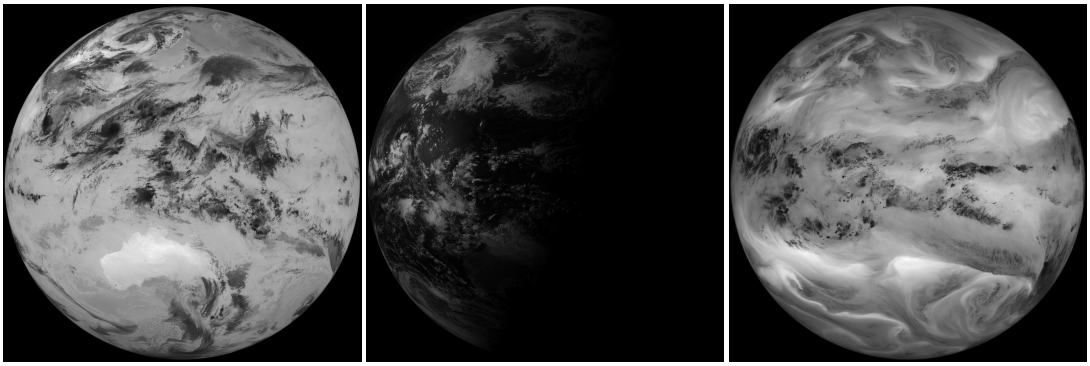


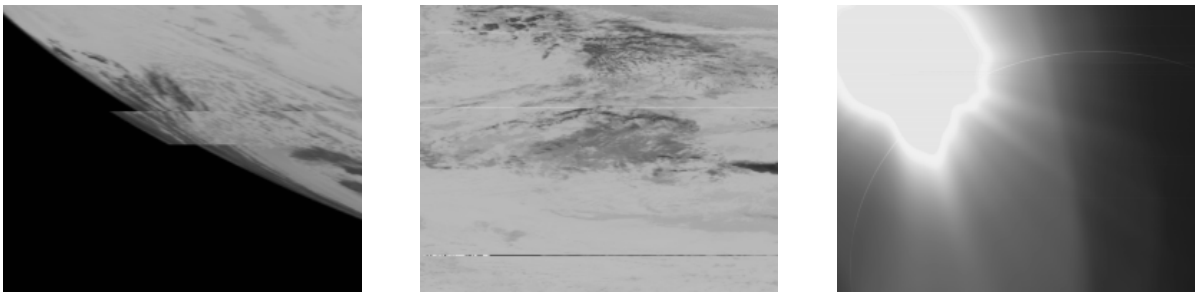
Figure 2.3: Nominal MTSAT-1R images across three spectral bands. From left to right: infrared (IR), visible (VIS), and water vapour (WV).

chance of sunlight falling into the imaging instruments directly. These anomalies can be detected reliably without use of the actual data using calibration methods described in Köpken [24]. As such, no automated method is needed to detect these anomalies.

Besides the origin of the anomaly, there is also the severity and classification level. There are 27 identified anomalies, classified among 7 detail levels [40]. The different detail levels will be explained below.

2.3.1. Image-Level Anomalies

Image-level anomalies affect the entire or large parts of an image. These anomalies indicate issues with the complete data acquisition or processing for a given time. This is often because of permanent failure in the satellite's instruments, data transmission, or other processing systems. Detection requires analysis of the full image like Earth disk fitting.



(a) Shifted scanlines error.

(b) Radiometric error.

(c) Straylight interference.

Figure 2.4: Different kinds of errors that can occur in satellite imagery.

2.3.2. Area-Level Anomalies

Area-level anomalies affect partial regions within the image that can be captured by one or multiple bounding boxes. These show local issues that impact bigger continuous sections of the data but do not extend across the entire image. They often result from temporary instrument malfunctions or data transmission errors affecting specific segments. Detection involves finding regions with abnormal values compared to the surrounding areas.

2.3.3. Scanline-Level Anomalies

Scanline-level anomalies affect one or more horizontal lines in the image. These anomalies show problems in the line-by-line scanning process, including detector readout issues, mirror stepping errors, or data transmission problems affecting individual scan lines. These anomalies are the most common anomalies found in the data. This is for example due to the sequential nature of spin-scan radiometer operation. Detection methods typically look at horizontal profiles, line-to-line differences, and metadata associated with individual scanlines.

2.3.4. Column-Level Anomalies

Column-level anomalies affect vertical lines or sets of vertical lines in the image. These are less common than scanline anomalies because the satellites scan from east to west and not north to south. When they occur, it often has to do with data buffer problems, memory failures, or issues in the data recording and not with the scanning instruments. Detection requires analysis of vertical profiles and column-to-column consistency.

2.3.5. Horizontal Pattern Anomalies

Horizontal pattern anomalies involve repeating patterns in the horizontal direction with set spacing. They are different from scanline anomalies in that they involve a repeating pattern. The repetition is typically linked to the number of detectors in the satellite. They are caused by calibration differences between individual detectors or variations in detector performance. Detection methods utilise the observed repetitions of these patterns through techniques like spectral analysis or pattern matching at known intervals.

2.3.6. Pixel-Level Anomalies

Pixel-level anomalies affect individual pixels or small groups of pixels. These are the smallest anomalies and often caused by smaller detector failures, or bit errors in data transmission. Unlike persistent detector issues that create regular patterns, pixel anomalies are mostly random. Detection methods focus on identifying outliers compared to neighbouring pixels using spatial filtering and sliding window analysis.

2.3.7. Synthetic anomalies

For the purpose of this research, synthetic anomalies are introduced to evaluate the model's ability to detect anomalies, these anomalies consist of anomalies that are similar to the real anomalies, and anomalies with the goal to pinpoint specific weaknesses of the model. These anomalies are broken scanline, black patches in the image, and black image around borders, celestial bodies reproduced as white circles, noise patches in the image and half swap, where the left and right half or top and bottom of the image are swapped. These anomalies are designed to be similar to real anomalies, but they are not found in the dataset, so they are not included in the definition of an anomaly[1]. Examples of these anomalies are found in Figure 2.5.

[1] More information on the construction of these synthetic anomalies can be found in Appendix A.

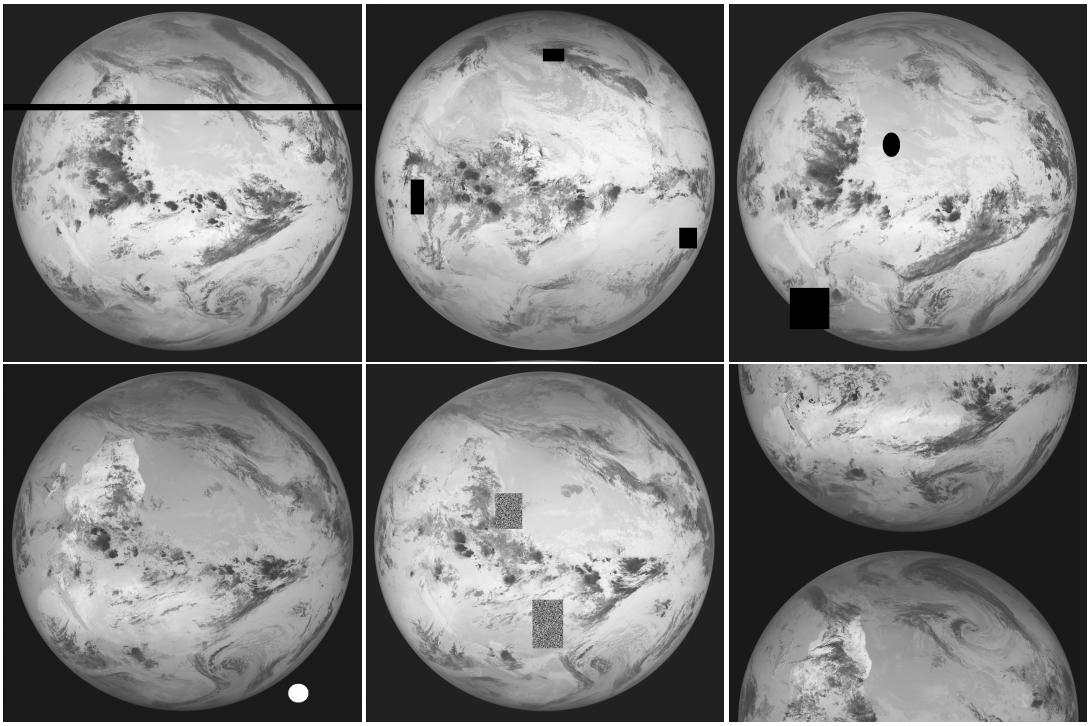


Figure 2.5: Examples of the six synthetic anomaly types used for evaluation. Top row, left to right: broken scanlines, black border patches, black patches. Bottom row, left to right: celestial body (white circle), noise patch, half swap.

2.4. Rule-based Anomaly Detection

Anomaly detection in geostationary satellite imagery is an active research field. Because researchers depend on high-quality data for analysis, robust noise and artifact detection algorithms are important. Traditionally, rule-based methods are used to find specific artifacts [23, 40]. These rule-based algorithms are effective, but require manual parameter tuning and input from domain experts to ensure reliable detection for different sensors and satellite types. This makes the method hard to scale, and adapt to newer satellites. Another weakness is that since only a small subset is examined for new anomalies, it is possible that rare anomalies are missed, and will not be detected automatically.

These traditional rule-based algorithms use variations of sliding-window statistics, local averaging, or other neighbourhood-based anomaly detection techniques.

Earlier stages of the GIAD project introduced a broader and more systematic detection framework. This includes Earth-disk fitting for identifying misalignment issues, hot/cold-pixel analysis, and metadata-based anomaly detection [40, 18]. More in-depth description of the GIAD rule-based detection algorithms can be found in Appendix B. With rule-based anomaly detection it is possible to identify anomalous images, and easily find the anomalous parts of the image. The rule-based algorithms, while effective, are limited in the ability to scale, and automate anomaly detection in new unseen data. This limitation raises the question if it is possible to use machine learning to automate the anomaly detection, to make it more scalable and adaptable to new datasets.

3

Related Work

Anomaly detection in images has been an active research area for over multiple decades. The field started with simple statistical methods, but with the rise of machine and deep-learning, and more computational power, newer techniques are able to learn the image structure and distribution, removing the need to make assumptions about the data distribution.

The earliest methods used in anomaly detection in images are statistical methods [44, 21, 30, 43]. Many of these methods are able to find anomalies based on some statistical property of the data, for this, most methods assume that the data follows a Gaussian distribution. When this assumption is not met, the performance of these methods can drop.

More modern methods are based on machine and deep learning. Machine learning approaches require low computational costs, and are able to do good anomaly detection on simple datasets [46]. These methods work by either learning what nominal data looks like, and then classifying data that falls outside the nominal learned data as anomalous [22, 16, 51]; or by learning to isolate anomalies, based on the intuition that anomalies are rare, and have different properties than nominal samples, making them sensitive to isolation [26]. These methods are not able to do anomaly localisation, because they use input vectors and no image representation.

Machine learning approaches are less suitable for anomaly detection in high dimensional data, like images [10]. The simplicity of the model makes it hard to find meaningful features of images, because they lack the ability to learn and store these in the model.

Deep learning methods, utilising neural networks to analyse and learn data representations, are able to learn more complex structural features of images. Autoencoders (AE) are a common architecture used for anomaly detection in images. Different variations like Classical Autoencoders [37, 47, 58], Variational Autoencoders (VAE) [38, 37, 4] and Deep Autoencoders [61, 14, 12] can be used for anomaly detection tasks. These studies show that the application of AEs for anomaly detection in images is a suitable and promising technique. AEs are trained to reconstruct nominal images from a learned compression, and compare the reconstruction to the original image; the anomalous images get a worse reconstruction than nominal images, enabling the detection.

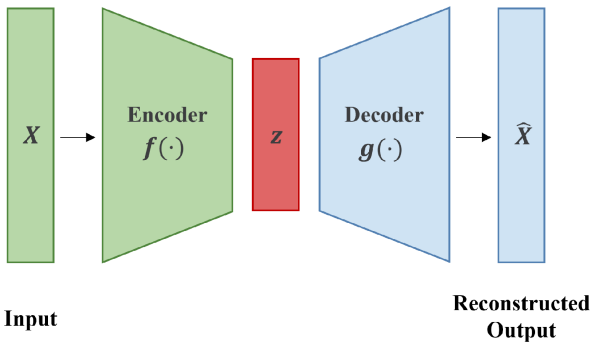


Figure 3.1: Overview of an autoencoder [20]

Another common anomaly detection method uses Generative Adversarial Networks (GANs). Originally, GANs are an unsupervised learning method to learn deep representations of specific data, using generative modelling [7, 15, 5]. GANs employ adversarial learning to train a generator, that learns to generate new images, and a discriminator, that judges the quality. The model learns by letting these models compete, and learn to outperform each other. An overview of the GAN architecture is shown in Figure 3.2.

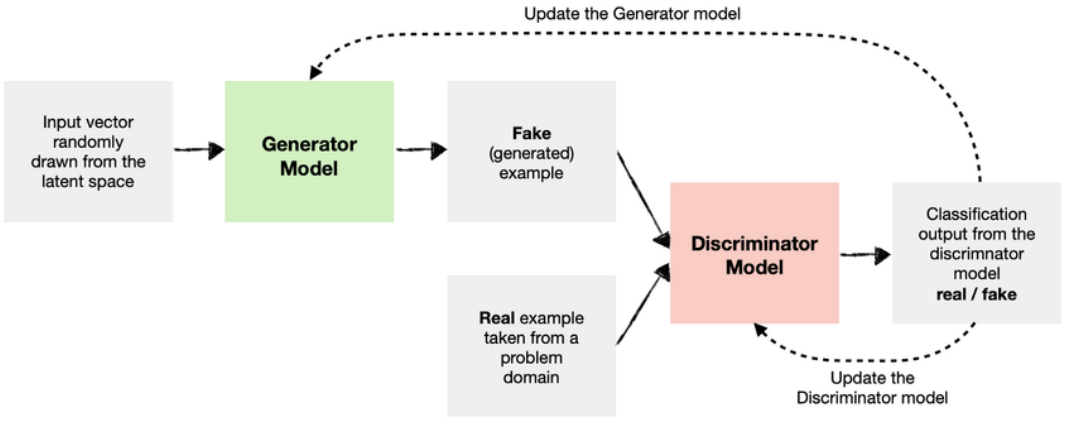


Figure 3.2: General architecture overview of a GAN [31]

When applied in anomaly detection in images, the networks used for the generator and the discriminator are CNNs. The GAN is trained on nominal data. This means that the generator will only approximate the distribution of nominal samples [49, 3]. Anomalies then can be detected by the discriminator, which recognises the anomalous images as fake. To be able to do localisation, the generator can be used to reconstruct the input image, this is done in methods like f-AnoGAN [48] and GANomaly [3]. In f-AnoGAN, the anomalies are scored and localised by using reconstruction error, and the critic's latent space difference between the input and the generated image. In GANomaly, the generator is an autoencoder, and the discriminator is a binary classifier that distinguishes between real and generated images. The anomaly score is calculated based on the reconstruction error of the generator and the output of the discriminator, making localisation less straightforward than in f-AnoGAN.

Diffusion models are a newer type of generative models [8]. These models have also been used for anomaly detection in images [53, 27, 28]. These models have shown to be able to create high quality reconstructions, and therefore good anomaly maps. Diffusion models learn to generate images by first slowly adding noise to an image, to then reverse that process and learn to denoise images step by step. This highly iterative process does enable the model to learn good data representations, but comes at the cost of high computational costs. Because this thesis focuses on developing a generalised anomaly detector suitable for real-world applications with big datasets, and limited computational resources, the high computational costs of diffusion models, when combined with the need for large datasets, make them less suitable for this research, and therefore they will not be further discussed in this thesis.

It is possible for deep learning models to learn to adapt to new unseen data, by using transfer learning. Transfer learning enables models to use an existing model, and retrain it to a small new dataset. Because the model already learned to find features, and recognise certain images, this knowledge can be fine-tuned to the new data set in a short time [63, 59]. This is often used in settings where little good training data is available.

Similar work has been done on anomaly detection in satellite imagery. A previous project applied a Variational Autoencoder (VAE) to exactly the same dataset of this thesis [52], training exclusively on anomaly-free satellite images so that anomalous inputs would produce higher reconstruction errors. Two preprocessing strategies were evaluated. In the downsampling approach, images were compressed to 150×150 pixels; the model reconstructed the overall earth disk well, but aggressive downsampling caused smaller anomalies to disappear before they reached the model. In the tiling approach, images were split into 150×150 tiles to preserve resolution, but this greatly increased the variance of the training distribution — individual tiles could contain open ocean, land, cloud, or disk boundaries — and the VAE failed to generalise across this variety. The study concluded that reconstruction-based detection is feasible when training variance is controlled, but that a larger latent capacity and a more expressive generative model are needed to handle the full diversity of the dataset. These findings highlight the limitations of VAEs, but suggest that the general reconstruction-based

approach for anomaly detection is promising. More complex models, with more latent capacity might be able to reliably detect anomalies.

An alternative approach for anomaly detection in satellite imagery was proposed by Esmaili et al. [11], which employed supervised transfer learning using pre-trained convolutional neural networks. The study developed binary classification models using ResNet-18 to detect anomalies in both the Advanced Baseline Imager (ABI) aboard GOES satellites and the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard polar-orbiting satellites. The ResNet-18 model achieved 94% accuracy for ABI anomaly detection with precision of 0.89, recall of 1.0, and F1 score of 0.94 on an independent evaluation dataset from GOES-18. This approach demonstrated that bigger models are able to learn complex features, and can find anomalies in the data. However, this supervised approach requires a large labelled dataset of both nominal and anomalous images, which does not align with the goal of a generalised anomaly detector.

The main different between the previous studies, and this thesis, is this thesis aims to investigate the potential ways to develop a generalised anomaly detector for satellite images. This generalisation needs to happen across multiple spectral bands, image sizes and bit depths.

For the model to be able to detect anomalies across different bands, the easiest solution is to create a model that is trained on images from all spectral bands. This way the model learns to reconstruct images from all these different bands. One such error is that sometimes, an image of one spectral band ends up being stored as another spectral band. To detect this error, the model needs to have specific information about how the different spectral bands look, and to what band they belong. For this, metadata augmentation can be used. Metadata can be used to augment the input data in different ways. The most straightforward is to encode the metadata in an array and append it to the input of the generator or discriminator, as proposed by [36]. In this approach, the metadata is added at the start of the model, and then the models should learn to use the metadata throughout its layers.

For different image sizes, there are a few options. The data can be normalised to a standard input size in the data pre-processing step of a model. This is a simple approach, but too much resizing can cause image details to be lost due to compression. Since the images used in this research can become large (22,000×22,000 pixels for Band 3 on the Himawari 9[1]) when compared to images from other sources, data compression and resizing the images is desirable to keep the anomaly detector running in reasonable time. Anomaly detection in images is often tested on low resolution images (32×32 - 512×512) [25, 6, 48]. For the biggest satellite images this would mean that the images have to be downsampled 40+ times. This means that a lot of data is lost in doing so. There are methods looking into anomaly detection for images with resolution ranging from 1k (1024×1024) [45, 9, 29] up to 4k (4096 × 4096) [60].

Most of these methods utilise some sort of tiling technique, where the images are split up in smaller tiles, then these tiles are used to train the detection methods, and then anomaly detection is performed. There are variations, where the tiles slightly overlap, to make sure context around the anomaly is fully considered. Although the tiling approach is a proven technique, for really high resolution images, a 1024×1024 tile would still mean that there is a total of 484 tiles per 22,000×22,000 images. A downside of the tiling approach is that anomalies that are bigger than a tile, or have to do with misalignment of the image, might not be detected, because the tile is reconstructed correctly, and the model cannot detect that the tile is in the incorrect place. To counteract this, metadata can be used to give the model information about where the tile is located in the image, and what the surrounding tiles look like. This way the model can learn to detect anomalies that are bigger than a tile, or misalignment of the image.

Lastly, the different satellites have different bit depths, varying between 8 and 12. This can be handled by normalising the bit depth in image pre-processing to values between 0 and 1. By doing this, the input for the model is always the same, and the model can handle the different bit-depths.

3.1. Research gap

As discussed above, anomaly detection in multi-band satellite imaging is an active research field. Many of the current approaches make use of methods that are specific to a specific instrument or satellite type. The current methods also mainly focus on detection of one, or a small group of anomalies. While there are some papers that focus on detection of multiple anomalies, they use different methods for these anomalies [23]. In current literature there is a lack of research in generalised anomaly detectors for a wider group of anomalies, as well as the issue that there is no literature investigating anomaly detectors that work for different satellites or measurement instruments. In addition, most anomaly detection research is benchmarked on

[1] https://www.data.jma.go.jp/mscweb/en/himawari89/space_segment/spsg_ahi.html

low-resolution images, and literature on anomaly detection in high-resolution imagery remains limited.

This thesis addresses the generalisation gap by:

1. Developing an unsupervised learning anomaly detection model, which does not require labelled anomalous data, and can identify novel anomalies.
2. Developing tiling strategies to scale to high-resolution imagery while preserving anomaly detail.
3. Evaluating cross-satellite and cross-band generalisation with controlled experiments.
4. Analysing the similarities and differences between the rule-based GIAD approach and the learned detector developed in this thesis, together with a recommendation on how the two approaches can support each other.
5. Investigating conditioning a model on satellite metadata to improve generalisation across instruments and bands.

These contributions relate to the research questions, where RQ 1 is related to contribution 2, RQ 2 relates to contribution 3 and 5, and RQ 3 is related to contribution 1 and 4.

4

Methods

The research gap identified in Chapter 3 can be closed by an unsupervised detector that scales to high-resolution multi-band satellite imagery, generalises across instruments, and requires no labelled anomaly examples. This chapter describes the methods proposed to address that gap. First, a convolutional autoencoder is implemented as a baseline, to establish the foundation for deep-learning-based anomaly detection methods, and provide a first comparison for the more complex f-AnoGAN framework. Then, the f-AnoGAN model is described in detail, including its architecture, training procedure, and anomaly scoring strategy. Finally, a metadata embedding approach is introduced to condition the model on satellite and channel information, and a transfer learning approach is proposed to evaluate the model's generalisation to new datasets.

4.1. Convolutional Autoencoder for Anomaly Detection

Before introducing f-AnoGAN, a convolutional autoencoder is implemented as a controlled baseline. It provides a reconstruction-error comparison without adversarial training, mode-collapse risk, or a separate encoder-inversion stage. Including it allows the additional complexity of the GAN framework to be isolated: any improvement of f-AnoGAN over the autoencoder is attributable to the richer generative model rather than to the scoring strategy.

4.1.1. Principle

As introduced in Section 3, an autoencoder learns to reconstruct images by compressing input data through an encoder, passing it through a bottleneck, and reconstructing the image from that bottleneck representation. The autoencoder is trained exclusively on nominal data. At inference time, the idea is that anomalous inputs produce higher reconstruction errors, since the bottleneck cannot represent patterns absent from the training distribution. This reconstruction error forms the basis of the anomaly score derived in Section 4.1.3.

4.1.2. Architecture

Encoder

Each encoder stage has a *multi-scale convolution block* with parallel 3×3 and 5×5 branches. Using two kernel sizes in parallel captures structure at different spatial scales within a single layer: the 3×3 branch responds to local pixel-level detail while the 5×5 branch captures broader texture patterns. This is relevant because satellite anomalies vary from single-pixel noise to multi-tile artefacts, and a single kernel size would bias the encoder towards one scale. The two branches are normalised independently with Group Normalisation before combining. Group Normalisation is used instead of Batch Normalisation because it does not depend on batch statistics, which vary when tiles from different satellites and spectral bands are mixed in the same batch. A fusing convolution with Group Normalisation and ReLU activation follows. Spatial downsampling is performed by a strided convolution rather than pooling, preserving more spatial information for reconstruction.

Bottleneck
The number of encoder stages L is chosen automatically so that the spatial resolution reaches a 16×16 bottleneck regardless of input size:

$$L = \left\lfloor \log_2 \frac{H}{16} \right\rfloor \tag{4.1}$$

Here, H is the height of the input tile (128 pixels in this case). For 128×128 tiles this yields $L = 3$ stages and a bottleneck of 16×16×256 units. The spatial compression forces the model to encode structural features rather than memorise pixel values; no skip connections are used, so anomalous patterns cannot bypass this compression at inference-time. A final multiscale block is applied at the bottleneck resolution to refine the representation before decoding.

Decoder
Each decoder stage mirrors its encoder counterpart. Bilinear upsampling followed by a convolution doubles the spatial resolution; this avoids the checkerboard artefacts that can be caused by transposed convolutions. The same 3×3/5×5 parallel-branch structure is retained. Instance Normalisation replaces Group Normalisation in the decoder, as it preserves per-image contrast rather than normalising across the channel group. This is relevant given the multi-satellite origin of the training data (MSG, GOES, Himawari), where radiometric contrast varies across sources, and accurate pixel-level reconstruction depends on retaining these per-image intensity characteristics. A final 1×1 convolution projects to a single output channel, and a Tanh activation maps pixel values to $[-1, 1]$.

No skip connections are used between encoder and decoder. This ensures that all information must pass through the bottleneck, preventing anomalous patterns from bypassing the compression at inference time.

4.1.3. Anomaly Scoring

Pixel-space error
The first anomaly score component for an image $\mathbf{x}$ is the mean squared error between the input image and its reconstruction, computed in the normalised $[-1, 1]$ space:

$$s_{\text{pixel}}(\mathbf{x}) = \frac{1}{HW} \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \tag{4.2}$$

Here W is the width of the image, and $\hat{\mathbf{x}}$ is the reconstructed image.

Feature-space error
Pixel-level MSE may be insensitive to structured anomalies that are spatially smooth or that the decoder partially reconstructs through generic texture synthesis. To address this, a feature-space error term is introduced, measuring the distance between the ℓ_2-normalised bottleneck representations of the input and of its own reconstruction:

$$s_{\text{feat}}(\mathbf{x}) = \left\| \frac{f_\phi(\mathbf{x})}{\|f_\phi(\mathbf{x})\|_2} - \frac{f_\phi(\hat{\mathbf{x}})}{\|f_\phi(\hat{\mathbf{x}})\|_2} \right\|^2 \tag{4.3}$$

where $f_\phi(\mathbf{x})$ and $f_\phi(\hat{\mathbf{x}})$ are the bottleneck representations of the input and its reconstruction respectively, each divided by their ℓ_2 norm to project them onto the unit sphere before taking the squared distance. This normalisation ensures the score is bounded regardless of the absolute magnitude of the bottleneck activations. If the reconstruction is faithful, encoding it again should yield a representation close to the original. An anomalous region that survives pixel reconstruction will nonetheless produce a different bottleneck code, since the encoder has only learned to represent nominal manifolds.

The combined anomaly score is:

$$s(\mathbf{x}) = s_{\text{pixel}}(\mathbf{x}) + \alpha \cdot s_{\text{feat}}(\mathbf{x}) \tag{4.4}$$

where s_{pixel} and s_{feat} are the pixel-space and feature-space errors defined above, and α is a weighting coefficient. $\alpha = 1.0$ throughout this work; equal weighting was not tuned further since the autoencoder serves as a baseline and the relative scaling of the two error terms is studied in the context of f-AnoGAN instead. Both score components are raw (unbounded) values. At evaluation time, s_{pixel} and s_{feat} are normalised jointly across the test set to $[0, 1]$ before combining, so that neither term dominates by scale (see Section 4.5). Note that this normalisation depends on the full test distribution and cannot be applied to a single image at deployment time; a calibration set would be required in a production setting. This formulation is analogous to the *izif* score used in f-AnoGAN [48], where an encoder maps images into the GAN latent space and feature-space distance supplements pixel-space reconstruction error.

4.1.4. Training

The model is trained on nominal satellite tiles of 128×128 pixels, created from pre-tiled 1024×1024 images. The full data preparation pipeline, including bit-depth normalisation, tiling parameters, and synthetic anomaly construction, is described in Appendix A. Tiles whose raw pixel values are entirely zero (i.e. outside the earth disk) are discarded before normalisation. All other tiles, including legitimately dark tiles such as night-side or low-radiance channels, are retained so that the model learns the full range of the nominal data distribution. Images are normalised to $[-1, 1]$ using the global bit-depth maximum (65 535 for 16-bit imagery) rather than per-tile statistics, to prevent bright-normalised background tiles from producing noisy input images.

Training uses the Adam optimiser with learning rate $\eta = 10^{-4}$, batch size 64, and a ReduceLROnPlateau scheduler that halves the learning rate when validation loss plateaus for 10 consecutive epochs. Mixed-precision training (FP16) reduces GPU memory consumption. Early stopping is applied with a patience of 250 epochs. MSE is used as the reconstruction loss. As the autoencoder solely serves as a baseline model, no extensive hyperparameter search was conducted.

4.2. f-AnoGAN

For the main anomaly detection model, the f-AnoGAN framework [48] is implemented. This method is chosen over the GANomaly [3], because of the more intuitive anomaly scoring and localisation strategy of f-AnoGAN, based on reconstruction error in both pixel and feature space.

4.2.1. Principle

f-AnoGAN detects anomalies by training a WGAN-SN and the encoder only on nominal data. At inference time, the model tries to reconstruct a test image through its learned nominal manifold. High reconstruction errors indicate an anomalous image.

The f-AnoGAN framework maps an input image to a latent vector via the encoder. The generator then reconstructs the image from that latent vector. The reconstructed image is evaluated by the critic, the WGAN term for what is called the discriminator in standard GAN literature. The critic produces an unbounded real-valued score using the Wasserstein objective; this allows the gradient signal to remain informative even when the generator improves. A higher critic score indicates a less realistic reconstruction, and thus a greater likelihood of anomaly.

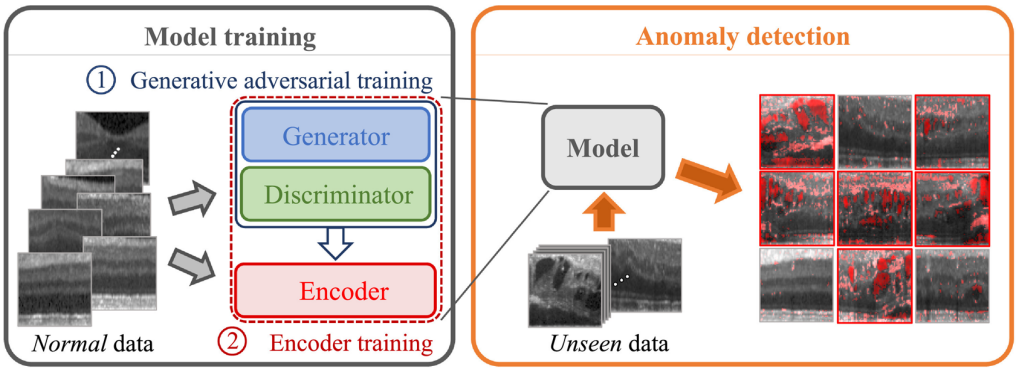


Figure 4.1: Overview of the f-AnoGAN framework [48]. Training consists of two stages: (1) adversarial training of the generator and discriminator on nominal data, followed by (2) encoder training to invert the frozen generator. At inference time, unseen images are encoded and reconstructed; high reconstruction error (shown in red) indicates anomalous regions.

4.2.2. Architecture

The framework consists of three components. The **generator** G learns to translate a d_z-dimensional Gaussian latent vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a 128×128 image via a CNN for upsampling: a linear projection to a 4×4 spatial map, followed by five nearest-neighbour upsample–convolution blocks with BatchNorm and ReLU, and a final Tanh activation. The **critic** D mirrors the generator with five strided-convolution blocks and Spectral Normalisation, collapsing the 128×128 input down to a scalar score. Unlike the original GAN, which trains the critic as a binary classifier, the Wasserstein objective measures how far the generated distribution is from

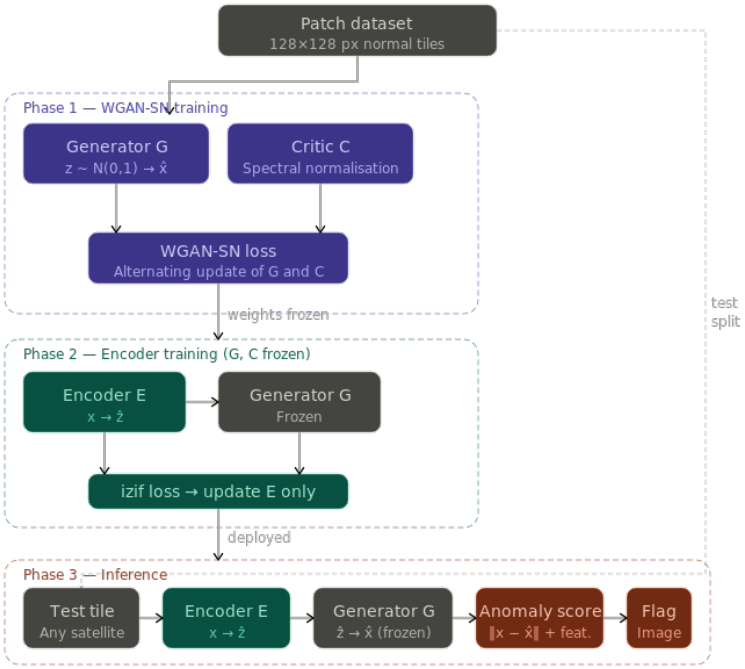


Figure 4.2: Full model pipeline used in this work. Satellite images are tiled and passed through the f-AnoGAN encoder–generator pair. The pixel-level reconstruction error and the feature-level discrepancy are combined into a per-tile anomaly score; the worst-tile score across the image is used as the image-level anomaly score.

the real one, yielding a more stable training signal; the Lipschitz constraint ensures this distance estimate gives meaningfull gradients, and is enforced here via Spectral Normalisation.

The **encoder** E maps a 128×128 image back to a d_z-dimensional latent vector, as proposed in Schlegl et al. [48]. It consists of five plain downsampling stages: each stage is a 4×4 strided convolution (stride 2, padding 1) followed by BatchNorm and LeakyReLU (slope 0.2), halving the spatial resolution and doubling the channel count up to a cap of 16×hidden_dim. No residual connections and no Spectral Normalisation are used in the encoder; these were found to constrain the inversion expressiveness without providing measurable benefit. The final 4×4 spatial map is flattened and projected to d_z via a linear layer.

4.2.3. Training

Training proceeds in two sequential phases.

Phase 1 – WGAN-SN. The generator and critic are trained adversarially on nominal tiles only. The generator tries to produce images that fool the critic; the critic tries to distinguish real from generated images. This adversarial dynamic drives both networks to improve jointly. Early stopping is not applied: in adversarial training, a drop in loss does not imply convergence, as the generator and critic continue to improve each other even at low loss values.

Phase 2 – Encoder (izif). Once the WGAN-SN is trained, its weights are frozen, since the encoder, has to learn to map images to the latent space of a specific Generator. Then the encoder is trained using an extended image-to-image-in-feature-space loss [48]:

$$\mathcal{L}_E(\mathbf{x}) = \mathcal{L}_{\text{img}}(\mathbf{x}) + \kappa \cdot \| f_c(\mathbf{x}) - f_c(G(E(\mathbf{x})))\|_2^2 + \lambda \cdot \mathcal{L}_{\text{reg}}(\mathbf{x}), \tag{4.5}$$

where E is the encoder, G the generator, f_c the critic feature extractor, κ the feature loss weight, λ the regularisation weight, and $\mathcal{L}_{\text{reg}}$ the latent regularisation term defined below. The image reconstruction loss is a 50/50 mix of MSE and L1:

$$\mathcal{L}_{\text{img}}(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x} - G(E(\mathbf{x}))\|_2^2 + \tfrac{1}{2}\|\mathbf{x} - G(E(\mathbf{x}))\|_1. \tag{4.6}$$

where $G(E(\mathbf{x}))$ is the reconstruction obtained by encoding $\mathbf{x}$ and decoding the resulting latent code. Pure MSE penalises large errors quadratically and produces blurry reconstructions; adding the L1 term showed better preservation of spatial sharpness in experiments, which is important for accurate anomaly localisation. The

second term compares raw (unnormalised) critic features f_c extracted at 40% network depth, with κ=3.0. Features are extracted at a shallow layer to retain spatial variance across inputs; deeper layers were found to collapse to near-constant feature vectors after ℓ_2 normalisation due to the Lipschitz constraint imposed by Spectral Normalisation.

The latent regularisation term $\mathscr{L}_{\text{reg}}$ penalises deviation of the encoded distribution from the $\mathcal{N}(\mathbf{0},\mathbf{I})$ prior that the generator was trained with, by matching the first two moments per latent dimension:

$$\mathscr{L}_{\text{reg}}(\mathbf{x}) = \underbrace{\frac{1}{d_z}\sum_{j=1}^{d_z}\bar{z}_j^2}_{\text{mean penalty}} + \underbrace{\frac{1}{d_z}\sum_{j=1}^{d_z}\left(\overline{z_j^2}-1\right)^2}_{\text{variance penalty}}, \tag{4.7}$$

where $\bar{z}_j$ and $\overline{z_j^2}$ are the batch mean and mean square of the j-th latent dimension, penalising deviations from $\mathbb{E}[z_j]$=0 and $\mathbb{E}[z_j^2]$=1. These penalties keep encoded representations aligned with the $\mathcal{N}(\mathbf{0},\mathbf{I})$ prior the generator was trained with. The weight λ=0.01 was found empirically; larger values degraded reconstruction quality.

The encoder is optimised with Adam (η=10^{-4}) and a ReduceLROnPlateau scheduler that halves the learning rate when the loss plateaus, with a minimum of 10^{-6}. The best-performing checkpoint (lowest mean training loss) is saved throughout training and restored as the final model.

4.2.4. Anomaly Scoring and Localisation

At inference time, the encoder maps the test image to a latent code and the generator reconstructs it. The anomaly score is:

$$s(\mathbf{x}) = \mathscr{L}_{\text{img}}(\mathbf{x}) + \kappa \cdot \| f_c(\mathbf{x}) - f_c(G(E(\mathbf{x})))\|_2^2, \tag{4.8}$$

where $\mathscr{L}_{\text{img}}$ is the image reconstruction loss (Equation 4.5), f_c the critic feature extractor at 40% network depth, and κ=3.0 the feature loss weight. The critic's scalar output is *not* included in the anomaly score; only image-space and feature-space reconstruction error are used. As with the autoencoder, both components are raw values that are normalised jointly across the test set to $[0,1]$ before combining at evaluation time (Section 4.5); this normalisation depends on the test distribution and cannot be applied to a single image at deployment without a calibration set. Full-resolution images are processed tile by tile without overlap; as discussed in Section 3, tile overlap is used in the literature to preserve context at tile boundaries, for the baseline experiments, overlap is not used, and overlap is introduced in a separate experiment to isolate its effect on performance.

4.3. Metadata embedding

As discussed in Chapter 2 and Section 3, the dataset contains images from different satellites and spectral bands. For each image, the satellite type, spectral band, and capture time are known. This metadata gives the model additional context about the image content. For example, water-vapour band (6–7 µm) show fundamentally different patterns than infrared window bands (10–13 µm), and images captured near local midnight have a higher chance of containing stray light artefacts from celestial bodies. Without metadata conditioning, the model must learn a single normal distribution over all channels and satellites simultaneously, which increases intra-class variance and may reduce anomaly detection sensitivity.

Following the projection discriminator approach of Miyato and Koyama [36], metadata conditioning is applied to all three components. The metadata vector is concatenated to the latent code $\mathbf{z}$ before the first linear layer of the Generator and before the final linear layer of the Encoder. The Critic is conditioned via a projection term that adds a dot product between the penultimate feature map and a learned metadata projection to the unconditional score, forcing the critic to assess whether the image is realistic *for its specific satellite and channel*, not merely in general. This approach is chosen because of the ease of implementation. The goal of also adding metadata embedding is to validate the potential of metadata embedding to the model, so implementing a more complex conditioning method, like cross attention, is out of scope for this research. The metadata vector is designed to capture the most relevant contextual information while remaining compact and learnable.

The metadata vector consists of four components:

Satellite class. Satellites are grouped into three classes: Meteosat (MSG series), MTSAT, and GOES. Each class is assigned a learnable embedding of dimension 4 via a `nn.Embedding` layer. The embedding is trained jointly with the generator and encoder. Grouping by class rather than individual satellite reduces sparsity while still capturing the systematic sensor differences between the three instrument generations.

Channel wavelength. The central wavelength of the spectral band is used as a single continuous scalar, normalised to $[0,1]$ over the full range of available channels (0.47–13.40 µm). Using the physical wavelength rather than a categorical embedding preserves the natural ordering of the spectrum: the model can interpolate between adjacent bands and generalise to unseen channel combinations.

Capture time. The sensing time is extracted from the image filename (format `YYYYMMDDHHMMSS`). Two cyclic encodings are derived to preserve the periodic structure of time:

$$\text{hour}: \quad \left(\sin\tfrac{2\pi h}{24}, \cos\tfrac{2\pi h}{24}\right) \tag{4.9}$$

$$\text{day-of-year}: \quad \left(\sin\tfrac{2\pi d}{365}, \cos\tfrac{2\pi d}{365}\right) \tag{4.10}$$

where h is the fractional hour and d is the day of year. Cyclic encoding ensures that midnight and 23:00 are treated as adjacent, avoiding the discontinuity that would arise from using raw hour values.

Tile spatial position. Each 128×128 tile is assigned its row and column index within the parent 1024×1024 image (indices 0–7 for an 8×8 grid). Both indices are normalised to $[0,1]$ and appended as two scalars:

$$\text{row\_norm} = \frac{r}{N_r - 1}, \qquad \text{col\_norm} = \frac{c}{N_c - 1} \tag{4.11}$$

where r, c are the zero-based tile row and column and $N_r, N_c = 8$ for 1024/128 tiling. This allows the model to learn position-dependent nominal patterns — for example, that the earth-disc limb tiles at the image boundary have systematically different appearance from central tiles, and that certain anomaly types (e.g. stray light) are more likely near the edges. The tile position is encoded directly in the tile filename (e.g. `_tile_r02_c03.png`) and parsed at training time.
The full metadata vector $\mathbf{m} \in \mathbb{R}^{11}$ is the concatenation of the satellite embedding (4), wavelength scalar (1), four time scalars (4), and two position scalars (2). It is appended to $\mathbf{z}$ so that the conditioned input becomes $[\mathbf{z}; \mathbf{m}] \in \mathbb{R}^{z_{\text{dim}}+11}$. Metadata conditioning is optional: when disabled, the model reduces exactly to the baseline f-AnoGAN, allowing a controlled comparison.

4.4. Transfer Learning

Transfer learning adapts a pre-trained f-AnoGAN to a new target satellite without training from scratch. Because the encoder is trained to invert a specific generator (the izif stage), any update to the generator's weights invalidates the encoder's mapping. The proposed approach therefore fine-tunes the two stages sequentially: the WGAN-SN is first fine-tuned on target-domain tiles at a reduced learning rate to shift the generated nominal manifold towards the new domain while preserving learned structure; the encoder is then re-trained from its pre-trained initialisation on the updated generator. The fine-tuning dataset size controls the trade-off between retraining time and performance.

4.5. Evaluation

Three models are evaluated: the convolutional autoencoder (Section 4.1), the f-AnoGAN model (Section 4.2), and the metadata-conditioned f-AnoGAN variant (Section 4.3). Each is assessed at two levels: image-level anomaly detection and pixel-level anomaly localisation. Pixel-level localisation is assessed qualitatively for all experiments.

4.5.1. Image-level detection

For each test image an anomaly score is computed as a weighted combination of pixel-space and feature-space reconstruction error (Section 4.2.4). Both score components are normalised jointly across the test set before combining, as described in Sections 4.1.3 and 4.2.4.

In addition to the mean score over all tiles of an image, two max-tile variants are reported:

- **Normalised max-tile**: the score of the worst-scoring tile after joint normalisation across the test set. Because normalisation is applied before taking the maximum, the value is comparable across experiments and reflects how anomalous the most suspicious tile looks relative to the full test distribution.
- **Absolute max-tile**: the raw worst-tile reconstruction error before normalisation. This value is scale-dependent and cannot be compared directly across experiments, but it is not affected by the distribution of the remaining tiles in the test set — making it more sensitive to localised anomalies that do not elevate the mean score.

Two metrics are reported:

- **AUROC**: area under the receiver operating characteristic curve, measuring separability between nominal and anomalous scores across all thresholds. AUROC is threshold-free and therefore independent of how the operating point is chosen.
- **F1 score**: harmonic mean of precision and recall, balancing the two equally:

$$F_1 = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.12}$$

  where precision = TP/(TP + FP) and recall = TP/(TP + FN). F1 is reported as a threshold-dependent summary metric; this is appropriate given that both false positives and false negatives carry operational cost in satellite monitoring. The decision threshold is swept over the percentiles of the combined score distribution; the threshold that maximises F_1 is selected and used to compute all threshold-dependent metrics.

4.5.2. Pixel-level localisation

The anomaly localisation maps (pixel error, feature error, combined) are visualised qualitatively for a representative set of test images. No pixel-level ground truth annotations are available, so localisation is not quantified and is evaluated qualitatively only.

4.5.3. Comparing rule-based GIAD and f-AnoGAN

Conducting a direct comparison between GIAD and f-AnoGAN is not straightforward, because the two methods are inherently different. GIAD is a deterministic, rule-based approach where a specific rule is defined to detect a given anomaly type. f-AnoGAN produces a continuous anomaly score, and a threshold must be chosen to decide whether an image is anomalous. This structural difference makes a quantitative comparison uninformative: GIAD has no threshold to vary, so there is no common operating point at which the two methods can be fairly evaluated against each other. Because of this, the two methods are compared qualitatively in the Discussion (Section 6.7), focusing on the specific applications of the methods, how they can enhance the user experience, where they overlap, and how they can support each other in practice.

5
Experiments

This chapter presents the experiments done to evaluate the f-AnoGAN approach and the autoencoder baseline. The experiments are structured to build on each other: the baseline is established first, then the following experiments each focus on a single proposed improvement — tiling strategy, spectral band subset, satellite domain, transfer learning, and metadata conditioning.

The baseline experiment (Section 5.1) evaluates both models on the full MTSAT/MSG dataset across all channels, establishing a performance reference for the remaining experiments. This experiment is related to RQ2 and RQ3 (See Section 1.2). The fine-grained anomaly detection experiment (Section 5.2) investigates the detection performance on pixel- and line-level anomalies, addressing RQ1. The tile overlap experiment (Section 5.3) investigates the effect of tile overlap on detection performance, addressing RQ1 and RQ3. The VIS-NIR and IR-WV experiments (Sections 5.4 and 5.5) evaluate the models on specific spectral band subsets, addressing RQ2. The transfer learning experiment (Section 5.6) investigates the ability of the model to generalise to unseen satellites, addressing RQ2. Finally, the metadata conditioning experiment (Section 5.7) investigates whether satellite and band metadata can improve detection performance, addressing RQ1.

For each of the three main experiments (baseline, overlap, and metadata conditioning), additional reconstruction and heatmap examples organised by classification outcome (TP, FP, TN, FN) are provided in Appendix C.

5.1. Baseline Experiments

The goal of this experiment is to establish a baseline performance of the f-AnoGAN implementation, giving a good starting point for the other experiments.

5.1.1. Dataset

The initial experiments are done on a dataset containing images from JMA MTSAT1, and EUMETSAT MSG9 satellites. The dataset contained all the different channels of both the satellites. Original images were first downsized to 1024×1024 images and were split up in 128×128 tiles, without overlap. This resulted in a dataset consisting of 189,100 training tiles, 21,012 validation tiles. The test set consisted of 1,599 nominal images and 1,771 anomalous images. The anomalous images are with synthetically introduced anomalies.

5.1.2. Setup

The different methods will be trained and evaluated on the same train, test and evaluation sets. Below the setup of the different methods are explained in more detail.

Autoencoder implementation

The autoencoder architecture explained in Section 4.1 is used for this experiment.

Table 5.1 lists the hyperparameters used during the training of the AutoEncoder.

f-AnoGAN implementation

The f-AnoGAN architecture explained in Section 4.2 is used for this experiment.

Table 5.2 lists the hyperparameters used during the training of the f-AnoGAN.

Table 5.1: Autoencoder training hyperparameters for the baseline experiment on MTSAT and MSG (all channels).

Hyperparameter	**Value**
Architecture	
Image size	128×128
Loss function	MSE
Training	
Epochs	100
Learning rate η	1×10^{-4}
Batch size	64
Optimiser	Adam
LR scheduler	ReduceLROnPlateau
LR decay factor	0.5
LR scheduler patience	10 epochs
Minimum learning rate	1×10^{-6}
Early stopping patience	20 epochs
Anomaly scoring	
Feature loss weight λ	1.0

Table 5.2: f-AnoGAN training hyperparameters for the baseline experiment on MTSAT and MSG (all channels).

Hyperparameter	**Value**
Architecture	
Image size	128×128
Latent dimension d_z	256
WGAN-SN training	
GAN epochs	50
Generator learning rate η_G	1×10^{-4}
Critic learning rate η_C	1×10^{-4}
Critic updates per generator step n_{crit}	1
Batch size	128
Encoder training (izif)	
Encoder epochs	20
Encoder learning rate η_E	1×10^{-4}
Feature loss weight κ	3.0
Latent regularisation weight λ	0.1

5.1.3. Results

The results of the methods are compared on the same test set, and the same evaluation metrics. The evaluation metrics are AUROC and the F1 scores. The results will be shown for the whole model, per satellite, per channel and per anomaly type. This gives a good insight of the overall model, on what parts it performs well, and where it performs worse. This is useful for understanding the model better, and for identifying where the model can be improved in the future.

f-AnoGAN

Table 5.3: Overall f-AnoGAN anomaly detection performance on the MTSAT/MSG baseline test set (1,599 nominal, 1,771 anomalous images). 95% bootstrap confidence intervals in parentheses.

Score component	AUROC	F1
Combined (pixel + feature)	0.6330 (0.6161–0.6503)	0.6948 (0.6805–0.7098)
Combined (max-tile normalised)	**0.7756 (0.7596–0.7910)**	**0.7256 (0.7086–0.7412)**
Combined (max-tile absolute)	0.6894 (0.6712–0.7074)	0.7037 (0.6899–0.7178)
Pixel mean	0.6561 (0.6379–0.6737)	0.6939 (0.6806–0.7081)
Pixel max-tile normalised	0.7652 (0.7485–0.7808)	0.7163 (0.7004–0.7322)
Pixel max-tile absolute	0.6637 (0.6459–0.6820)	0.6979 (0.6838–0.7126)
Feature mean	*0.5960 (0.5782–0.6146)*	*0.6931 (0.6799–0.7075)*
Feature max-tile normalised	0.7442 (0.7272–0.7605)	0.7126 (0.6973–0.7279)
Feature max-tile absolute	0.6631 (0.6445–0.6812)	0.6970 (0.6827–0.7117)

Table 5.3 shows the overall performance of the f-AnoGAN on the baseline test set. The combined max-tile absolute score achieves the highest AUROC of 0.6894 among the absolute and mean-based scores, outperforming all individual absolute max-tile component scores. The absolute pixel and feature max-tile scores (AUROC 0.6637 and 0.6631) improve modestly over the pixel mean (AUROC 0.6561) but fall below the combined max-tile. The mean feature score (AUROC 0.5960) is the weakest component, and the combined mean score (AUROC 0.6330) falls between the pixel and feature means. The normalised max-tile scores outperform all absolute and mean variants: pixel (0.7652), feature (0.7442), and combined (0.7756) each exceed their absolute and mean counterparts. These normalised max-tile scores will therefore be used for the rest of the experiments.

Figure 5.1 shows the score distribution of the combined max-tile score for nominal and anomalous images. The distributions overlap considerably, consistent with the moderate AUROC values.

Table 5.4: Per-anomaly-type f-AnoGAN detection on the baseline test set. Each type is evaluated against all 1,599 nominal images (random-classifier AUROC = 0.5). Absolute max-tile scores; best value per column is **bold**.

Anomaly type	AUROC-px-max	AUROC-ft-max	AUROC-cb-max
Noise	**0.9353**	0.5769	**0.8206**
Lines	0.7602	0.7206	0.6829
Patches	0.8334	**0.7840**	0.7620
Border black patches	0.7909	0.7300	0.6911
Celestial body	0.7226	0.6569	0.6550
Half swap	0.5083	0.5005	0.4945

When breaking down the performance per anomaly type as shown in table 5.4, it shows that line and noise anomalies are detected more accurately than the other anomalies. The half swap anomalies are the hardest to detect, where even a random classifier would perform better than the f-AnoGAN.

Table 5.5 shows that MET09 outperforms MTSAT1 across all score components. The differences are small, meaning that the model is able to learn the nominal patterns of both satellites equally well.

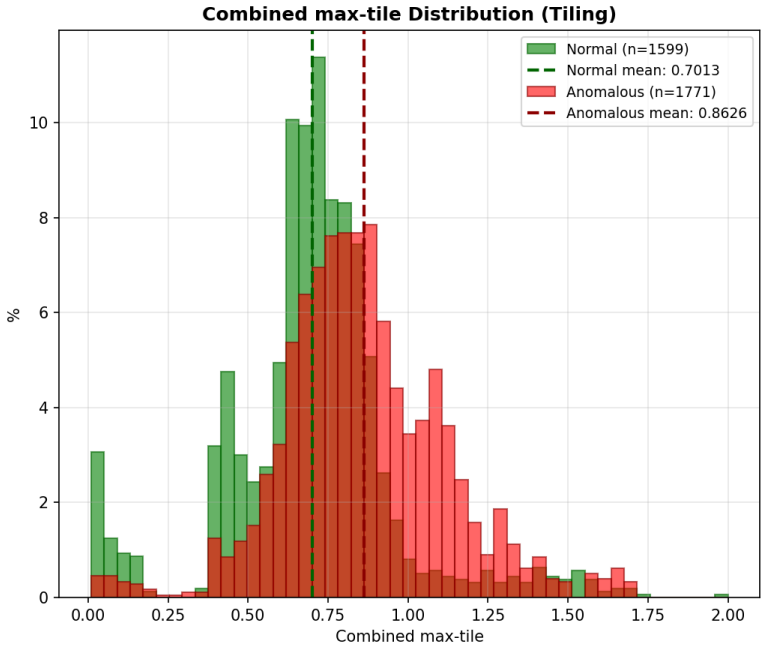


Figure 5.1: Score distribution of the combined max-tile anomaly score for nominal and anomalous images on the baseline test set. The distributions overlap considerably, consistent with the moderate AUROC values.

Table 5.5: Per-satellite anomaly detection on the baseline test set.

Satellite	AUROC-px-max	AUROC-ft-max	n_{anom}
MET09	0.7640	**0.7650**	1207
MTSAT1	**0.7699**	0.7154	564

Table 5.6: Per-channel pixel-level anomaly detection on the baseline test set. Best AUROC-px per satellite is **bold**; worst is *italic*.

Satellite	Channel	AUROC-px-max	AUROC-ft-max	n_{anom}
MET09	WV0620	**0.9183**	**0.8988**	114
MET09	WV0730	0.8901	0.8782	116
MET09	IR1340	0.8497	0.8606	116
MET09	IR0970	0.8577	0.8690	115
MET09	IR1080	0.7802	0.8333	113
MET09	IR1200	0.8270	0.8358	114
MET09	IR0870	0.8281	0.7907	116
MET09	IR0160	0.6912	0.6839	104
MET09	IR0390	0.6221	0.6446	114
MET09	VIS0080	*0.6352*	0.6569	99
MET09	VIS0060	0.6954	*0.6304*	86
MTSAT1	WV0670	0.8783	**0.8318**	114
MTSAT1	IR1200	**0.8869**	0.7925	113
MTSAT1	IR1080	0.8214	0.7782	114
MTSAT1	IR0375	*0.6524*	*0.5713*	115
MTSAT1	VIS0072	0.6553	0.7076	108

When looking at the performance of the different channels in Table 5.6, the water vapour channels consistently achieve the highest pixel AUROC (WV0620: 0.9183 for MET09, WV0670: 0.8783 for MTSAT1), while visible light and near-infrared channels score lowest. Interestingly, these near-infrared and visible light channels are affected by the day–night cycle of the Earth.

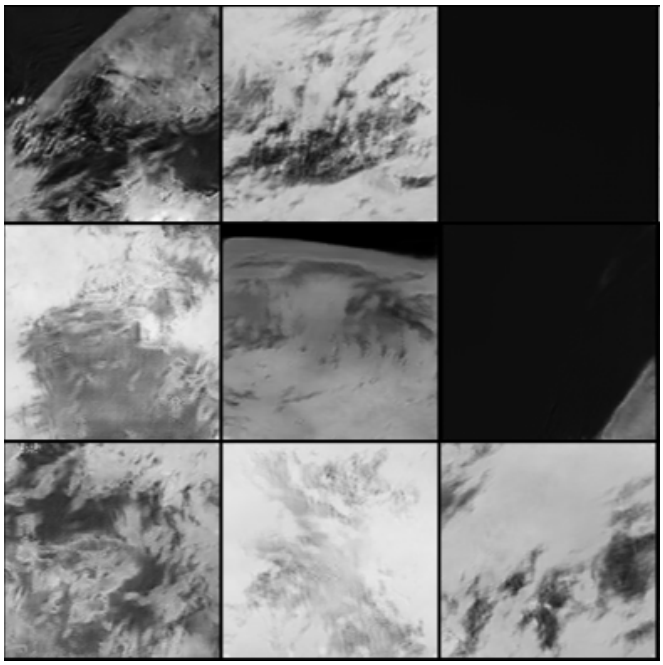


Figure 5.2: Random samples drawn from the trained baseline WGAN generator ($G(\mathbf{z})$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$). The variety of cloud structures, earth-limb positions, and spectral textures across tiles indicates that the generator has learned a diverse nominal manifold covering both MTSAT-1 and MSG-9 imagery.

Figure 5.2 shows random samples drawn from the trained baseline WGAN generator. The variety of cloud structures, earth-limb positions, and spectral textures across tiles indicates that the generator has learned to recreate different kinds of tiles present in the training set.

Figure 5.3 shows that the model responds very differently to normal and anomalous inputs. On the normal image, reconstruction error is lower and quite evenly spead accross the complete image, and no single tile stands out in the max-tile score map, consistent with the image being nominal. On the anomalous image, which contains a horizontal line artefact, the pixel error map shows a sharp bright band exactly along the line, and the corresponding row of tiles receives the highest max-tile scores. The reconstruction does not reproduce the line and the error map spikes precisely where the anomaly lies. This confirms that the model's scoring mechanism correctly localises this anomaly type.

Autoencoder

Table 5.7 shows the overall performance. The pixel max-tile score (AUROC = 0.7165) is the strongest aggregation, consistent with the same finding for f-AnoGAN. The autoencoder scores 0.059 AUROC points below the f-AnoGAN combined max-tile (0.7756), showing that the two-stage GAN approach offers an improvement over a plain reconstruction baseline.

Table 5.8 breaks down performance by anomaly type. Noise is easily detected (AUROC = 0.9562), which is consistent with the f-AnoGAN result; reconstruction error is a strong signal when a tile is uniformly corrupted. This score is even outperforming the f-AnoGAN pixel max-tile score for noise (0.9353), suggesting that the autoencoder's reconstruction error is particularly effective for this anomaly type. All other anomaly types score between 0.68 and 0.72, indicating that the autoencoder generalises less discriminatively across anomaly categories than f-AnoGAN does. Half-swap anomalies again fall below chance (0.4604), as expected for a tile-level detector with no spatial context.

The per-channel breakdown in Table 5.9 reveals both similarities and differences with f-AnoGAN. MET09 WV0620 is again the strongest channel (0.9745), and visible-light channels remain the weakest. The WV0620 outperforms the f-AnoGAN pixel max-tile score for this channel (0.9183), however when looking at the other channels, the autoencoder performs worse than the f-AnoGAN pixel max-tile score for all other channels.

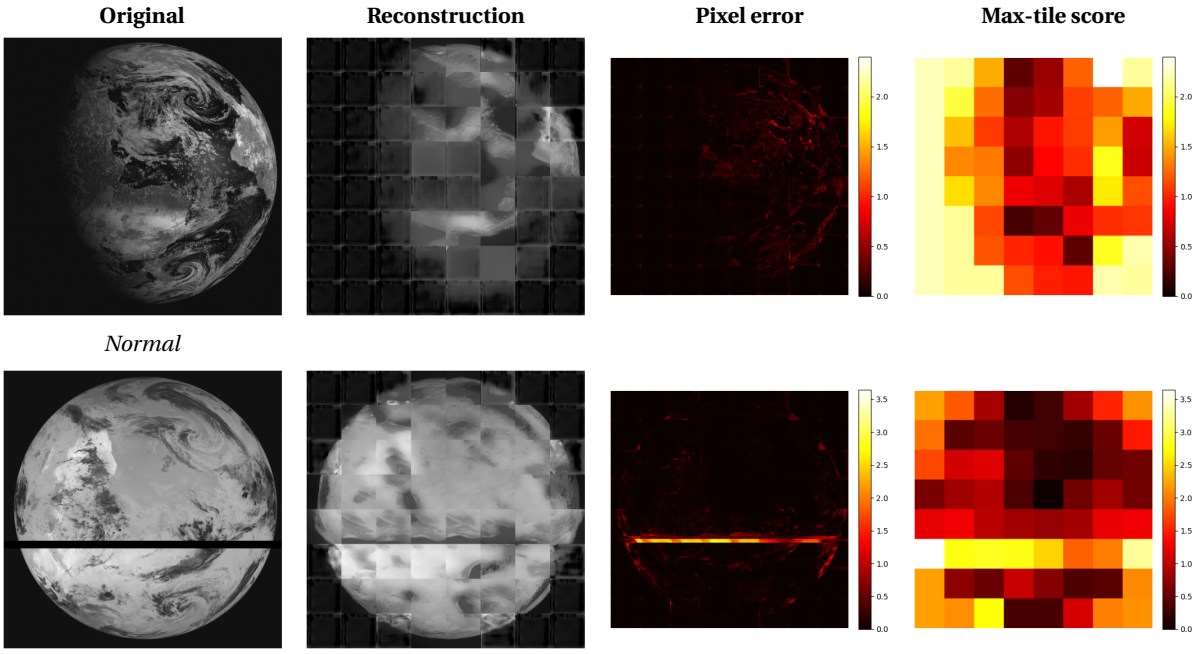


Figure 5.3: Reconstruction and anomaly scoring for a normal (top row) and an anomalous (bottom row) test image. Each row shows: the original full-resolution image, the encoder-driven reconstruction $G(E(\mathbf{x}))$, the pixel-level error map, and the max-tile anomaly score heatmap. For the normal image, error is low and diffuse across tiles. For the anomalous image, which contains a horizontal line artefact, the pixel error concentrates sharply along the line and the corresponding tile row receives the highest max-tile scores, demonstrating that the model correctly localises this anomaly type. Please note the difference in the colour scales of the error maps: the normal image's pixel maps have a lower max scale, casuing them to appear brighter than the anomalous image's pixel maps, which have a higher max scale.

Table 5.7: Overall autoencoder anomaly detection performance on the MTSAT/MSG baseline test set (1,599 nominal, 1,771 anomalous images).

Score component	AUROC
Combined max-tile	0.7165
Combined mean	0.6226
Pixel max-tile	0.7165
Pixel mean	0.6579
Feature max-tile	0.6383
Feature mean	0.5730
f-AnoGAN combined max-tile (reference)	*0.7756*

Table 5.8: Per-anomaly-type autoencoder detection on the baseline test set. Each type is evaluated against all 1,599 nominal images. Best value per column is **bold**.

Anomaly type	AUROC-px-max
Noise	**0.9562**
Patches	0.7222
Celestial body	0.7050
Lines	0.7033
Border black patches	0.6807
Half swap	*0.4604*

Table 5.9: Per-channel autoencoder pixel-level anomaly detection on the baseline test set. Best AUROC-px per satellite is **bold**; worst is *italic*.

Satellite	Channel	AUROC-px-max
MET09	WV0620	**0.9745**
MET09	WV0730	0.8856
MET09	IR1340	0.7748
MET09	IR1080	0.7681
MET09	IR0870	0.7671
MET09	IR1200	0.7618
MET09	IR0970	0.7475
MET09	IR0160	0.6923
MET09	IR0390	0.5955
MET09	VIS0080	0.5209
MET09	VIS0060	*0.4983*
MTSAT1	IR1080	**0.8186**
MTSAT1	IR1200	0.8161
MTSAT1	IR0375	0.6187
MTSAT1	VIS0072	0.5471
MTSAT1	WV0670	*0.5370*

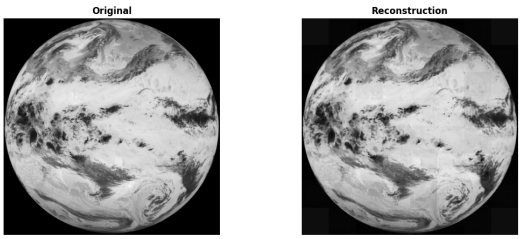


Figure 5.4: Reconstruction example from the autoencoder baseline model on the test set. The left panel shows the input image, and the right panel shows the autoencoder reconstruction.

However, the channel ranking diverges notably for MTSAT1: WV0670, which was the strongest MTSAT1 channel for f-AnoGAN (0.8783), drops to last place (0.5370) for the autoencoder. Conversely, MTSAT1 IR1080 improves from 0.8214 under f-AnoGAN to 0.8186 — nearly the same — while MTSAT1 IR1200 stays consistent (0.8869 vs. 0.8161). This suggests that the encoder–decoder pair of the autoencoder and the two-stage f-AnoGAN latent space capture different spectral features, especially in near-infrared channels.

Figure 5.4 shows a reconstruction example from the autoencoder baseline model on the test set. The reconstruction is of higher quality when compared to the f-AnoGAN reconstruction. However, this does not mean that the autoencoder is better at detecting anomalies, as this means that the anomalies get reconstructed as well, since the autoencoder generalises too well. This is shown in figure 5.5, where the nominal and anomalous score distributions overlap more than in the f-AnoGAN case (Figure 5.1).

5.2. Fine-grained anomalies

One of the main questions of the thesis is how well automatic anomaly detection methods can capture and recognise a wide range of anomalies, including fine-grained anomalies on the pixel level of an image. In the second experiment, this will be investigated. The goal of this experiment is to see how well the anomaly detection method is able to capture and recognise anomalies that consist of only a few pixels. For this experiment, the model trained for the initial experiment will be used. A separate test set is created for this experiment.

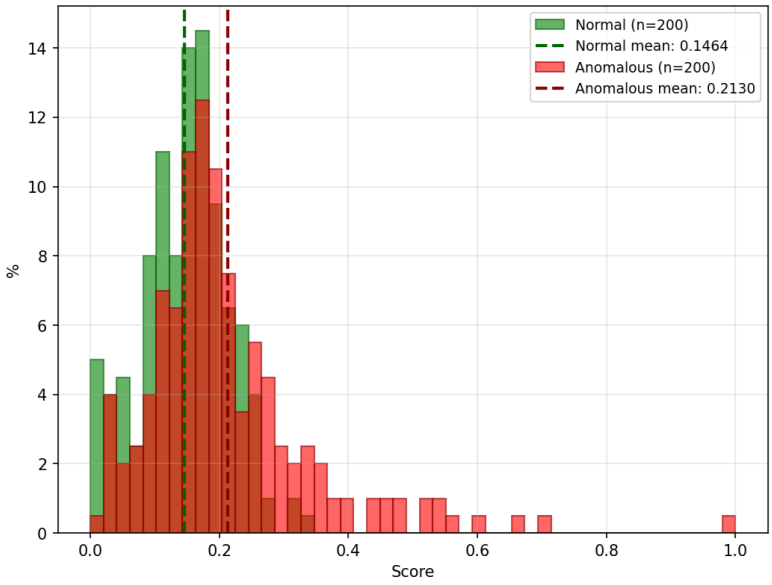


Figure 5.5: Score distribution of the autoencoder pixel max-tile anomaly score for nominal and anomalous images on the baseline test set. The distributions overlap considerably, consistent with the moderate AUROC values.

This test set contains images with fine-grained anomalies. Keeping the results of the baseline experiment in mind, it is expected that the model will not perform well on the fine-grained anomalies. This is because in the baseline experiment, the model already struggled to capture all the bigger anomalies.

5.2.1. Dataset

The fine-grained anomalies are synthetically added to the full resolution images. Then the images will be downsized to 1024×1024 and tiled to 128×128 tiles. This way, the experiment captures how well the model can handle downsampled fine-grained anomalies, which is the setting in which the model will be used in practice. The fine-grained anomalies mainly consist of individual broken scanlines, broken columns, or patches of only a maximum of 1% of the total pixels in the image.

5.2.2. Setup

The exact f-AnoGAN model from the baseline experiment is used, the parameters are shown in Table 5.2

5.2.3. Results

Table 5.10: Overall f-AnoGAN anomaly detection performance on the fine-grained test set (1,599 nominal, 1,699 anomalous images). Normalised max-tile takes the per-image maximum after jointly scaling tile scores to $[0,1]$; absolute max-tile takes the maximum over raw reconstruction errors. 95% bootstrap confidence intervals in parentheses.

Score component	AUROC	F1
Combined max-tile (normalised)	0.5145 (0.4945–0.5339)	0.6814 (0.6672–0.6966)
Pixel max-tile (normalised)	0.5285 (0.5092–0.5489)	0.6815 (0.6675–0.6966)
Feature max-tile (normalised)	0.5045 (0.4846–0.5242)	0.6810 (0.6671–0.6960)
Combined max-tile (absolute)	0.6053 (0.5869–0.6253)	0.6856 (0.6715–0.7005)
Pixel max-tile (absolute)	0.5966 (0.5772–0.6159)	0.6861 (0.6715–0.7014)
Feature max-tile (absolute)	0.5768 (0.5567–0.5966)	0.6832 (0.6687–0.6982)
Baseline combined max-tile (no overlap, separate model)	*0.7756*	*(0.7596–0.7910)*

The overall performance of the f-AnoGAN on the fine-grained test set is shown in Table 5.10. The normalised max-tile score is close to random performance, indicating that the model cannot reliably detect fine-

grained anomalies when averaging error across the image. In contrast to the baseline experiment, the absolute max tile metrics did perform better than the normalised max tile errors. This might be caused by the fact that the smaller anomalies do not raise the mean error of a tile that much, but does show a spike in absolute error in the raw error maps.

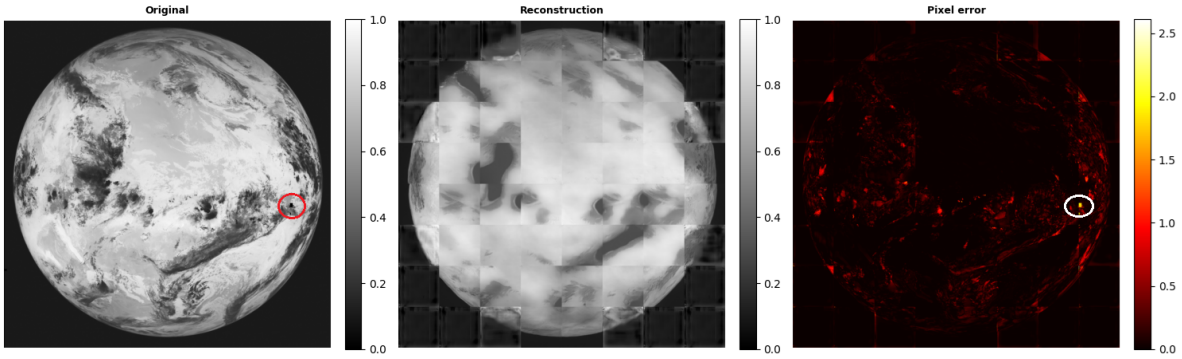


Figure 5.6: Reconstruction example from the baseline model on a fine-grained anomalous image. Left: input image; Middle: model reconstruction; Right: pixel-level error map. The fine-grained anomaly produces only a localised spike in the absolute error map, which is not sufficient to raise the mean tile error above the nominal threshold. The error is highlighted with a red circle on the input image, to indicate the location of the anomaly.

Figure 5.6 shows an example of the reconstruction and pixel error map from the baseline model on a fine-grained anomalous image. The anomaly is a small patch of black pixels, which produces only a localised spike in the absolute error map. This spike is not sufficient to raise the mean tile error above the nominal threshold, which explains why the normalised max-tile score performs close to random performance. The absolute max-tile score does perform better, but still struggles to reliably detect these fine-grained anomalies.

Table 5.11: Per-anomaly-type f-AnoGAN detection on the *fine-grained* test set. Each type is evaluated against the same 1,599 nominal images as the baseline (random-classifier AUROC = 0.5). Absolute max-tile scores; best value per column is **bold**.

Anomaly type	AUROC-px-max	AUROC-ft-max	AUROC-cb-max	n
Lines	0.6111	0.6618	0.6545	319
Celestial body	0.6069	0.5638	0.6072	319
Patches	0.5688	0.6488	0.6099	253
Noise	**0.7367**	0.5226	**0.6956**	319
Border black patches	0.5385	0.5637	0.5475	233

For the different anomalies, we see that again the noise anomaly is the easiest to detect, with an AUROC of 0.7367, which is quite good. The other anomaly types have an AUROC that is close to 0.6 or lower, which means that they are detected only slightly better than random performance. The pattern of the best recognised anomalies is the same for the fine-grained anomalies as for the nominal anomalies, despite the different metric used.

Table 5.12: Per-satellite pixel-level anomaly detection on the fine-grained test set. Both satellites perform near the random baseline (AUROC-px $\approx$ 0.5).

Satellite	AUROC-px-max	AUROC-ft-max	n_{nominal}	n_{anom}
MET09	**0.6101**	**0.5933**	1100	1276
MTSAT1	0.5714	0.5530	499	575

Here it can be seen that for the fine-grained anomalies, the MTSAT1 has a worse performance than the MET09 satellite, this might be because the MTSAT1 images are a bit darker than the MET09 images, and that

most of the anomalies involve black patches or lines, making them harder to detect, especially when the anomalies are only a few pixels big.

Table 5.13: Per-channel pixel-level anomaly detection on the fine-grained test set. Best AUROC-px per satellite is **bold**; worst is *italic*. For this experiments, for all channels, except for MTSAT VIS0072, there where 100 nominal and 116 anomalous images. For the MTSAT VIS0072 channel, there were 99 nominal and 111 anomalous images.

Satellite	Channel	AUROC-px-max	AUROC-ft-max		
MET09	IR1340	0.7331	0.6954	100	116
MET09	WV0620	**0.9010**	**0.8310**		
MET09	IR0970	0.7229	0.7314		
MET09	IR1200	0.6145	0.6504		
MET09	WV0730	0.7949	0.7234		
MET09	IR1080	0.6108	0.5831		
MET09	IR0390	0.5115	0.5484		
MET09	IR0870	0.6044	0.5524		
MET09	VIS0060	0.6144	0.5723		
MET09	VIS0080	*0.5187*	0.5141		
MET09	IR0160	0.5695	*0.4809*		
MTSAT1	WV0670	0.5299	**0.6093**		
MTSAT1	IR1200	**0.6318**	0.5499		
MTSAT1	IR1080	0.5967	0.5588		
MTSAT1	VIS0072	0.5758	0.5424		
MTSAT1	IR0375	*0.5102*	*0.5183*		

For the different channels, shown in Table 5.13, the higher wavelength channels have better performance when compared to the lower wavelength channels. This is similar as the baseline results.

5.3. Tile overlap

In the baseline experiment, the images were split into non-overlapping tiles. The results of this experiment showed that full image reconstructions suffered from tile border artifacts, because the independent generated tiles did not have smooth transitions between them. The aim of this experiment is to see if using overlapping tiles can reduce the tile border artifact. The idea is that by having some overlap between the tiles, the model can learn to generate smoother transitions, and the overlapping parts of the tiles can be smoothed together to reduce the border artifacts.

5.3.1. Dataset

The dataset for this experiment will mostly be the same images and satellites as in the baseline experiment, however, now the tiles will have 16 pixel overlap on each side, this means that on the 1024x1024 images, with 128×128 tiles and 16 pixel overlap, meaning that instead of 8 tiles per row and column, there are 9 tiles per row and column, resulting in 81 tiles per image instead of 64. The overlapping parts of the tiles will be averaged together to create the final image reconstructions.

5.3.2. Setup

The f-AnoGAN architecture is the same as in the baseline experiment, as shown in Table 5.2, with the same hyperparameters. The only difference is that there are now more tiles per full image, and that these tiles have overlap, and this need to be stitched together when looking at the reconstruction. How this is done, does not influence the scoring, because that is done over the complete individual tiles, aggregated together afterwards.

5.3.3. Results

Two stitching modes were evaluated on a model trained from scratch under the overlap experiment setup (same architecture and hyperparameters as the baseline, but trained as a separate run): *average*, where overlapping tile regions are blended by taking the mean, and *crop*, where each pixel is taken from the first tile that covers it (first-write-wins). Both modes use a 16-pixel overlap on a 128×128 tile grid, producing 81 tiles per image instead of 64. Note that this is an indicative comparison, as the two models were trained independently on separate runs. A controlled comparison using the original baseline model evaluated with overlap is described in Section 5.3.4.

Table 5.14: Overall f-AnoGAN anomaly detection performance with overlapping tiles (overlap=16) on the full test set (1,599 nominal, 1,771 anomalous images). The model was trained separately from the baseline under the same hyperparameter configuration. 95% bootstrap confidence intervals in parentheses.

Score component	AUROC	95% CI
Combined max-tile	0.8155	(0.8002–0.8292)
Pixel max-tile	0.8258	(0.8105–0.8397)
Feature max-tile	0.7619	(0.7456–0.7774)
Baseline combined max-tile (no overlap, separate model)	*0.7756*	*(0.7596–0.7910)*

The results in Table 5.14 show that the f-AnoGAN trained with overlapping tiles achieves a combined max-tile AUROC of 0.8155, which is a substantial improvement over the baseline model's 0.7756 (trained without overlap). The pixel max-tile score performs best with an AUROC of 0.8258, while the feature max-tile score is lower at 0.7619, but still shows improvement over the baseline feature max-tile (0.7442). These results suggest that using overlapping tiles during training can significantly enhance anomaly detection performance, likely by reducing tile border artifacts and providing more contextual information to the model.

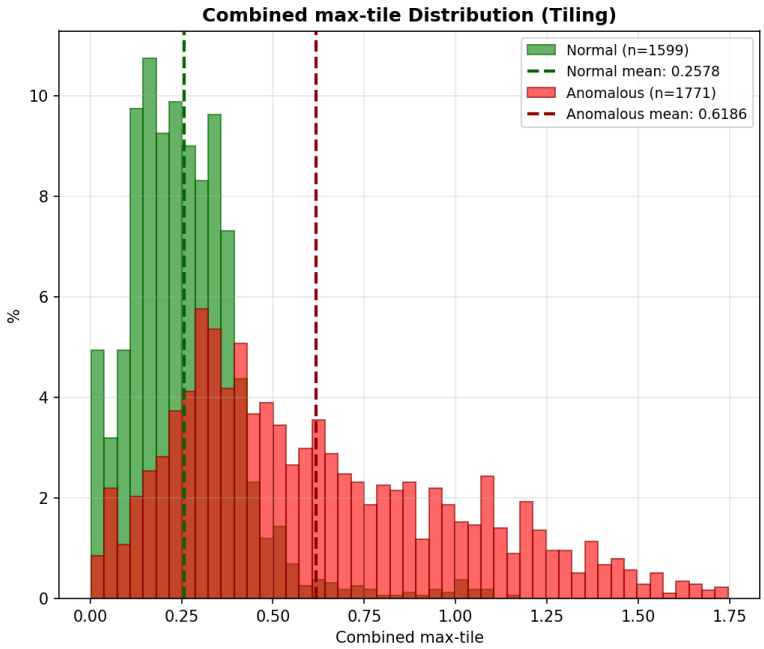


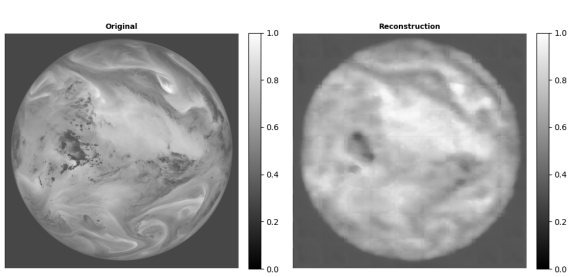
Figure 5.7: Score distribution of the combined max-tile anomaly score for the overlap experiment. Compared to the baseline (Figure 5.1), the nominal and anomalous distributions show greater separation, consistent with the improved AUROC.

The score distribution in Figure 5.7 shows that the nominal and anomalous distributions are more separated compared to the baseline model (Figure 5.1), which is consistent with the improved AUROC. The nominal scores are generally lower, while the anomalous scores are higher, indicating better discrimination between nominal and anomalous images.

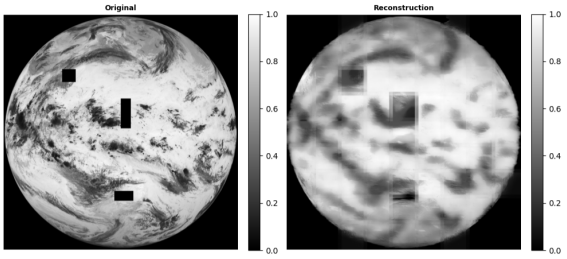
Table 5.15: Per-anomaly-type detection with overlapping tiles (average stitching). Results for crop stitching are identical to four decimal places. Each type evaluated against all 1,599 nominal images. Best value per column is **bold**.

Anomaly type	px-max	ft-max	cb-max	n
Noise	**0.9786**	**0.9857**	**0.9858**	319
Lines	0.7825	0.7561	0.7773	311
Patches	0.8476	0.7893	0.8329	284
Border black patches	0.8072	0.7795	0.8082	282
Half swap	0.7942	0.5795	0.7139	256
Celestial body	0.7378	0.6499	0.7066	319

Again, the noise anomaly has the highest score and again, the pixel error is often the best determinator for anomalies. Notably, the half swap combined max-tile AUROC is 0.7139 (Table 5.15), compared to 0.4945 in the baseline (Table 5.4), a substantial improvement for an anomaly type that was previously below random. The addition of extra global information of the image, by having more tiles, and some duplicate information in the training process, might explain this improvement in performance.



(a) Normal image. The reconstruction closely follows the original, with noticeably smoother tile boundaries than the non-overlapping baseline.



(b) Anomalous image containing rectangular block artefacts.

Figure 5.8: Reconstruction examples from the overlap model on a normal (left) and an anomalous (right) test image. Each panel shows the original (left) and encoder-driven reconstruction $G(E(\mathbf{x}))$ (right). Overlapping tiles with average stitching produce smoother transitions than the non-overlapping baseline.

Figure 5.8 shows that the overlapping tiles with average stitching produce smoother transitions at tile boundaries compared to the baseline model's non-overlapping reconstructions. On the anomalous image, the block artefacts are absent from the reconstruction, as the encoder projects the input onto the nominal manifold; this mismatch between original and reconstruction is what drives the anomaly score at the affected tiles.

5.3.4. Overlap evaluation on the baseline model

For the overlap experiment executed above, a separate dataset, with a specific overlapping tiles was used to train the model. However, it is also interesting to see how well the overlapping strategy works, when only used during evaluation. For this, the baseline model, trained on non-overlapping tiles is used, and during evaluation, the test images are tiled with overlap. With this, the influence of training and evaluation can be shown.

Table 5.16: Controlled overlap evaluation: original baseline model evaluated with 16-pixel overlap (average stitching) vs. no overlap. 1,599 nominal, 1,771 anomalous images.

Evaluation mode	AUROC-comb-max-tile	AUROC-px-max-tile	AUROC-ft-max-tile
No overlap (baseline)	0.7756 (0.7596–0.7910)	0.7652 (0.7485–0.7808)	0.7442 (0.7272–0.7605)
Overlap=16, average	0.6759 (0.6583–0.6944)	0.6603 (0.6418–0.6793)	0.6270 (0.6067–0.6458)

The controlled evaluation shows that applying 16-pixel overlap to the original baseline model at evaluation time does not improve performance — the combined max-tile AUROC drops from 0.776 to 0.676. This shows the importance of having training and evaluation datasets that use the same tiling approach.

5.4. Visible and Near Infrared band experiment

The results of experiment 5.1 showed varying performance across the different anomalies, spectral bands and small differences between the satellites. One of the biggest differences in performance was observed between Near Infrared (NIR) and Visible (VIS) bands, that had a significantly lower performance than the other bands. A possible reason for this is the day night cycle that is visible in these bands, is underrepresented in the training data, and thus not learned properly. To investigate this hypothesis, this experiment will focus on training and evaluating a model trained only on these VIS and NIR bands, to see if a separate model can learn this day-night cycle.

5.4.1. Dataset

The dataset will consist of only images from the VIS and NIR bands (From MET09: VIS0060, VIS0080 and IR0160; From MTSAT1: VIS0072), from the same satellites as in the baseline experiment. The images will be tiled in the same way as in the baseline experiment, with non-overlapping tiles of 128×128 pixels.

5.4.2. Setup

The f-AnoGAN architecture is the same as in the baseline experiment, with the same hyperparameters as shown in Table 5.2.

5.4.3. Results

For these exact settings, the f-AnoGAN model suffered from mode collapse, and was not able to learn a good representation of the data. A possible cause for this is that 40% of the dataset is black tiles. The low availability of good training images with actual earth disc and cloud patterns present, makes it hard for the generator to learn a good representation of the data, and for the discriminator to learn to distinguish between real and fake images.

Since a VIS/NIR-only model is not viable, the next experiment asks the reverse question: does removing the VIS and NIR bands from the full training set improve detection performance for the remaining channels, which are not affected by the day-night cycle?

5.5. IR and WV band experiment

The results of experiment 5.1 showed varying performance across the different anomalies, spectral bands and small differences between the satellites. There we saw that the performance of the NIR and VIS bands was significantly lower than the other bands, and thus we created a separate model for these bands. The question now is how much the performance of the baseline model would improve when it is trained only on the IR and WV bands that are not affected by the day-night cycle. This experiment will focus on training and evaluating a model trained only on these IR and WV bands, to see if the performance of the baseline model can be improved by removing the underrepresented day-night cycle from the training data.

5.5.1. Dataset

The dataset will consist of only images from the WV and TIR bands, from the same satellites as in the baseline experiment. The images will be tiled in the same way as in the baseline experiment, with non-overlapping tiles of 128×128 pixels.

5.5.2. Setup

The f-AnoGAN architecture is the same as in the baseline experiment, with the same hyperparameters as shown in Table 5.2.

5.5.3. Results

The model trained on only the WV and TIR bands performs only slightly better, indicating that the difference between a model trained with or without the VIS and NIR bands does not have that much impact on the total performance.

Table 5.17: Overall f-AnoGAN anomaly detection performance on the IR/WV-only test set (1,200 nominal, 1,374 anomalous images). 95% bootstrap confidence intervals in parentheses.

Score component	AUROC	95% CI
Combined max-tile	0.7962	(0.7795–0.8127)
Pixel max-tile	0.7800	(0.7624–0.7969)
Feature max-tile	0.7853	(0.7684–0.8028)
Baseline combined max-tile (all bands)	*0.7756*	*(0.7596–0.7910)*

Table 5.18: Score distributions for the IR/WV experiment: mean and standard deviation of anomaly scores for nominal and anomalous images (TIR and WV bands only).

Score component	μ_{nominal}	σ_{nominal}	μ_{anomaly}	σ_{anomaly}
Pixel max-tile	0.755	0.566	1.676	1.002
Feature max-tile	0.005	0.003	0.010	0.006
Combined max-tile	0.296	0.219	0.696	0.411

When looking at the score distributions it is interesting to see that the nominal scores are lower, with similar variance as the baseline. The anomalous pixel error is higher than the baseline model, this could mean that anomalies in the VIS and NIR bands on average had a low pixel reconstruction error.

Table 5.19: Per-anomaly-type f-AnoGAN detection on the IR/WV test set. Each type is evaluated against all 1,200 nominal images (random-classifier AUROC = 0.5). Best value per column is **bold**.

Anomaly type	AUROC-px-max	AUROC-ft-max	AUROC-cb-max	n
Noise	**0.9393**	0.7855	**0.9167**	240
Patches	0.8797	**0.9493**	**0.9198**	234
Lines	0.8541	0.9414	0.9036	239
Border black patches	0.8099	0.9103	0.8708	229
Celestial body	0.6403	0.5818	0.6213	240
Half swap	0.5060	0.4959	0.4995	192

The per-anomaly breakdown in Table 5.19 shows the same pattern as the baseline: lines and noise are the strongest performers, while half swap remains at random-classifier level (AUROC = 0.50) regardless of whether VIS/NIR bands are included.

The per-channel results in Table 5.21 confirm the pattern seen in the baseline: water vapour channels consistently achieve the highest AUROC-px for both satellites. With VIS and NIR channels removed, the worst-performing channels are now the short-wave infrared bands (IR0390 for MET09 and IR0375 for MT-SAT1), which still involve more complex, high-contrast cloud patterns than the thermal and water vapour bands.

5.6. Transfer Learning

Another question this research aims to answer is how well the anomaly detection model generalises to new satellites. In the baseline experiment, the model was trained and evaluated on two satellites (MTSAT-1 and MSG-9). This experiment will test how well, and with how much data the model can adapt to new unseen data. The model is expected to improve when new data is injected from a similar domain. This experiment tests how well the baseline model generalises to GOES-11, a satellite unseen during training. Four fine-tuning conditions are evaluated, including a zero-shot baseline.

Table 5.20: Per-satellite pixel-level anomaly detection on the IR/WV test set.

Satellite	AUROC-px-max	AUROC-ft-max	n_{nominal}	n_{anom}
MET09	**0.8314**	**0.8398**	800	918
MTSAT1	0.7029	0.7515	400	456

Table 5.21: Per-channel pixel-level anomaly detection on the IR/WV test set. Best AUROC-px per satellite is **bold**; worst is *italic*.

Satellite	Channel	AUROC-px-max	AUROC-ft-max	n_{nominal}	n_{anom}
MET09	WV0620	**0.9104**	0.8978	100	114
MET09	IR1340	0.8721	0.8538	100	116
MET09	IR0970	0.8649	**0.8669**	100	115
MET09	WV0730	0.8688	0.8872	100	116
MET09	IR1200	0.8776	0.8700	100	114
MET09	IR1080	0.8271	0.8325	100	113
MET09	IR0870	0.8269	0.8119	100	116
MET09	IR0390	*0.6719*	*0.7786*	100	114
MTSAT1	WV0670	**0.7970**	0.7628	100	114
MTSAT1	IR1200	0.7842	**0.8928**	100	113
MTSAT1	IR1080	0.6829	0.7945	100	114
MTSAT1	IR0375	*0.6470*	*0.6286*	100	115

5.6.1. Dataset

Source model

The baseline f-AnoGAN model from Section 5.1 is used as the starting point. It was trained on 189,197 tiles from MTSAT-1 and MSG-9 across all available spectral bands.

Fine-tuning data

Fine-tuning tiles are drawn from a pre-tiled GOES-11 dataset (128×128 tiles from 1024×1024 images). GOES-11 has 5 spectral bands. To ensure each band is represented equally, fine-tuning tiles are sampled per band.

Table 5.22: Transfer learning fine-tuning conditions. Images are sampled uniformly per spectral band from the GOES-11 training set. Each 1024×1024 image yields 32 non-empty 128×128 tiles.

Condition	Images per band	Total tiles
Zero-shot	0	0
Few-shot S	10	1 600
Few-shot L	100	16 000

Test set

The anomalous images are a mix of real anomalous images and synthetically introduced anomalies. The GOES11 images, by default have a different ratio than the MET9 and MTSAT1 images, where these images normally are square, the GOES11 images have only half the rows, compared to the width, while covering the whole earth disk. This means the GOES11 images are not downsampled to 1024×1024, but to 1024×512.

5.6.2. Setup

For the zero-shot condition, the baseline model is evaluated on GOES-11 without any weight updates. For the fine-tuning conditions, both the WGAN-SN and the encoder are fine-tuned on the sampled GOES-11 tiles. The WGAN-SN is fine-tuned for 50 epochs at a reduced learning rate of $\eta = 1 \times 10^{-5}$ (10× lower than baseline) to avoid overwriting the learned nominal manifold. The encoder is subsequently fine-tuned for 40 epochs at $\eta = 1 \times 10^{-4}$. All other hyperparameters match the baseline experiment (Table 5.2).

As an upper-bound reference, a model is also trained from scratch on GOES-11 data only, using the same architecture and hyperparameters as the baseline experiment (150 WGAN epochs, 150 encoder epochs; Table 5.2). This standalone model is trained on all available GOES-11 training tiles and evaluated on the same GOES-11 test set as the transfer learning conditions, providing a direct comparison between transfer learning and in-domain training.

5.6.3. Results

Table 5.23 summarises detection performance across all conditions. Figure 5.10 shows the progression of AUROC with fine-tuning budget alongside the standalone GOES-11 reference.

Table 5.23: f-AnoGAN anomaly detection on GOES-11 across all transfer learning conditions (500 nominal, 547 anomalous test images). Max-tile AUROC reported throughout; 95% bootstrap confidence intervals in parentheses. †Few-shot L uses mean-tile AUROC (max-tile evaluation was not available for this condition).

Condition	Tiles	Combined AUROC	Pixel AUROC	Feature AUROC
Zero-shot	0	0.618 (0.581–0.653)	0.587 (0.551–0.623)	0.642 (0.606–0.673)
Few-shot S	1,600	0.655 (0.619–0.686)	0.642 (0.606–0.673)	0.656 (0.619–0.686)
Few-shot L†	16,000	0.687 (0.651–0.717)	0.638 (0.601–0.670)	0.719 (0.685–0.748)
Standalone	14,832	0.696 (0.662–0.725)	0.692 (0.658–0.721)	0.707 (0.673–0.736)
Baseline combined max-tile	*0.7756*	*(0.7596–0.7910)*		

Zero-shot performance

When the baseline model is applied directly to GOES-11 without any fine-tuning (Table 5.23), the combined max-tile AUROC of 0.618 is above random, but the pixel score (0.587) is only slightly above chance, suggesting the pre-trained nominal manifold only partially transfers to the unseen satellite. The feature component (0.642) carries most of the discrimination, reflecting that the critic's latent space retains some useful structure even without fine-tuning. The score is nonetheless lower than the baseline model's performance on MTSAT/MSG (0.776), showing that domain shift between satellites substantially degrades zero-shot anomaly detection performance.

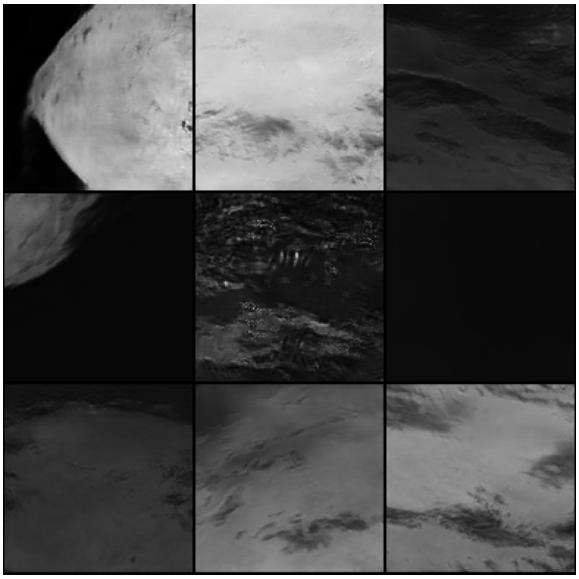
Few-shot S: 10 images per band

After fine-tuning on just 10 images per spectral band (1,600 tiles total), the combined max-tile AUROC rises from 0.618 to 0.655, with both pixel and feature components improving in roughly equal measure. This shows that even a minimal fine-tuning budget provides a meaningful gain.

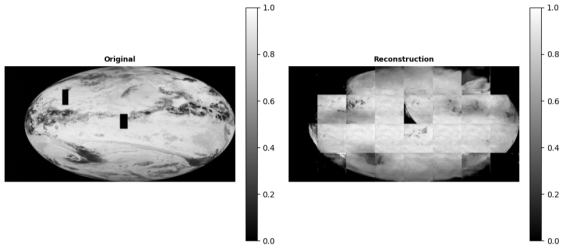
Few-shot L: 100 images per band

After fine-tuning on 100 images per spectral band (16,000 tiles), the combined AUROC improves further to 0.687. However, the 100-images-per-band condition does not outperform the standalone model (combined 0.696). Only the feature score (0.719) exceeds the standalone feature AUROC of 0.707; the pixel component (0.638) is substantially lower than the standalone pixel AUROC of 0.692.

Figure 5.9 illustrates the asymmetry between the generator and encoder after fine-tuning: the generator produces plausible GOES-11 tiles, but the encoder reconstruction does not accurately capture cloud patterns or the earth disc, indicating that the encoder has not successfully learned to invert the updated generator. When compared to the baseline generator samples (Figure 5.2), the retrained generator shows no real signs of performance decline, and keeps producing recognisable satellite tiles, however the encoder's failure to adapt likely explains the lower performance of the transfer learning.



(a) Random samples from the fine-tuned WGAN generator ($G(\mathbf{z})$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) after few-shot L fine-tuning. The tiles show recognisable GOES-11 structures, indicating the generator has adapted to the new domain.



(b) Encoder-driven reconstruction ($G(E(\mathbf{x}))$) from the few-shot L model on an anomalous GOES-11 test image. The reconstruction fails to capture cloud patterns or the earth disc, showing the encoder has not successfully learned to invert the updated generator.

Figure 5.9: Generator samples (left) and encoder reconstruction (right) from the few-shot L transfer learning model. The contrast illustrates that the WGAN successfully adapts to GOES-11 imagery, while the encoder fails to produce meaningful latent representations for the new satellite.

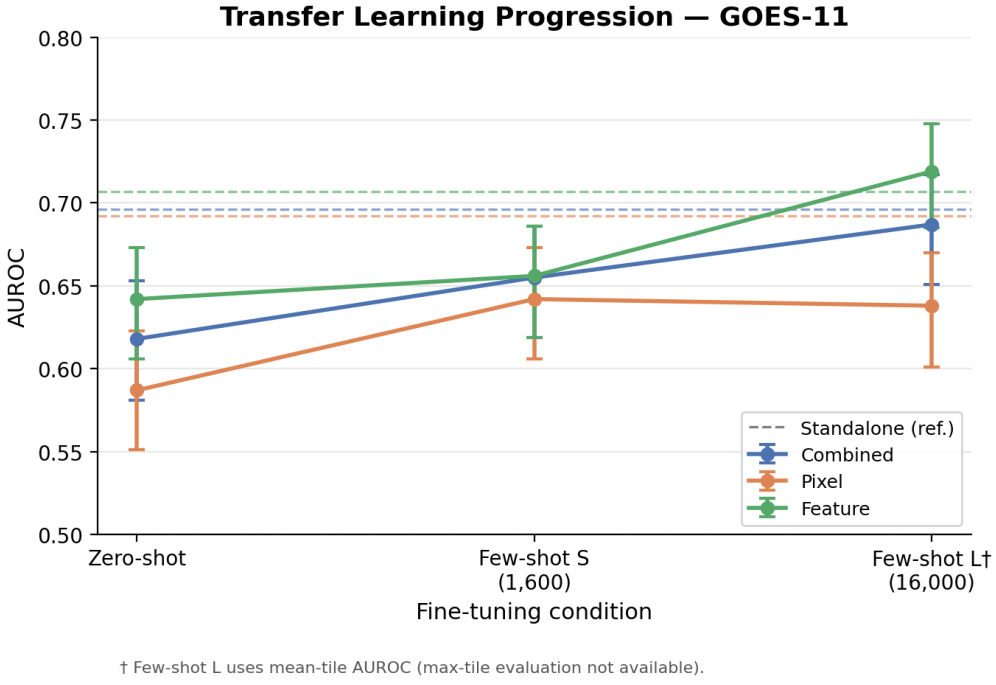


Figure 5.10: Transfer learning AUROC progression on GOES-11 as a function of fine-tuning budget. Error bars show 95% bootstrap confidence intervals. The dashed lines mark the GOES-11 standalone model as an upper-bound reference. Few-shot L uses mean-tile AUROC (†).

Figure 5.10 shows the progression of AUROC with fine-tuning budget alongside the standalone GOES-11 reference. Despite the progression, the standalone model trained from scratch on GOES-11 data outperforms all transfer learning conditions, suggesting that the pre-trained MSG/MTSAT model does not provide a beneficial initialisation for GOES-11, and that the conservative fine-tuning learning rate may prevent sufficient adaptation to the target domain.

GOES-11 standalone

The standalone model, trained from scratch on 14,832 GOES-11 tiles (≈93 images per channel across 5 channels), achieves a combined max-tile AUROC of 0.696, which exceeds all transfer learning conditions. For a consistent cross-condition comparison on mean-tile AUROC, the progression is: zero-shot 0.540, few-shot S 0.621, few-shot L 0.631, standalone 0.666. Notably, the standalone model uses roughly the same amount of data as few-shot L (≈93 images per channel vs. 100 per channel), yet outperforms it. This suggests that the pre-trained MSG/MTSAT model does not provide a beneficial initialisation for GOES-11, and that the conservative fine-tuning learning rate may prevent sufficient adaptation to the target domain.

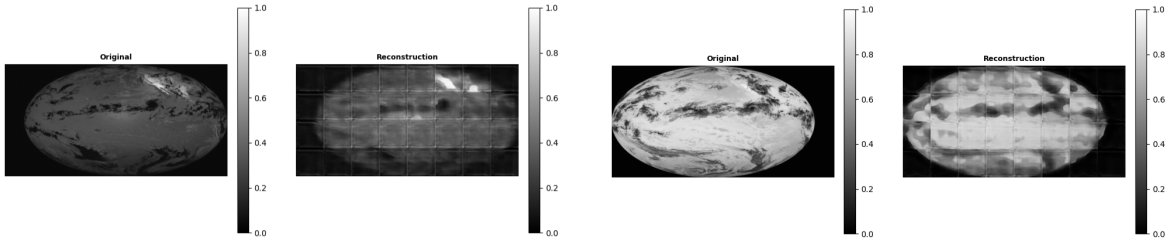


Figure 5.11: Two representative reconstructions from the standalone GOES-11 model on normal test images. Each pair shows the original (left) and the reconstruction $G(E(\mathbf{x}))$ (right). While tile-boundary artefacts remain visible, the reconstructions broadly preserve the earth-disc shape and large-scale cloud patterns, in contrast to the transfer learning model (Figure 5.9).

When comparing a reconstruction from the standalone GOES-11 model to the few-shot L model (Figure 5.11), the standalone model's reconstruction captures more of the cloud patterns and earth disc, indicating that it has successfully learned to invert its generator, which likely contributes to the slightly better anomaly detection performance compared to the transfer learning conditions.

5.7. Metadata conditioning

In the baseline results, the model was not able to capture some anomalies. One of these anomalies was the half swap anomaly, in which either the right and left side, or the top and bottom half of the image are swapped. The base implementation, does not have any information about the positions of the tile in the image. One way to capture this anomaly better is to provide the model with additional metadata to the tile. This experiment, will investigate the effect of metadata conditioning on the performance of the model. The expected outcome is that anomalies like the half swap will have a higher performance. Another possibility is that the extra data provided to the data, only makes it harder to find good image representations in the latent space, meaning the performance might drop.

5.7.1. Dataset

The same MTSAT/MSG dataset as the baseline experiment is used. Each tile is augmented with the corresponding metadata, parsed from the file path and image timestamp.

5.7.2. Setup

The metadata conditioning follows the design described in Section 4.3: an 11-dimensional vector comprising a 4-dimensional satellite-family embedding, channel wavelength, cyclic time-of-day and day-of-year encodings, and normalised tile position, injected into the generator, encoder, and critic. All other hyperparameters are identical to the baseline experiment (Table 5.2).

Table 5.24: Overall f-AnoGAN anomaly detection performance with metadata conditioning on the MTSAT/MSG test set (1,599 nominal, 1,771 anomalous images). 95% bootstrap confidence intervals in parentheses.

Score component	AUROC	95% CI
Combined max-tile	0.8071	(0.7911–0.8217)
Pixel max-tile	0.8016	(0.7856–0.8158)
Feature max-tile	0.7425	(0.7251–0.7585)
Baseline combined max-tile (no overlap, separate model)	*0.7756*	*(0.7596–0.7910)*

5.7.3. Results

The overall performance of the metadata-augmented model is shown in Table 5.24. The combined max-tile AUROC of 0.8071 exceeds the baseline AUROC of 0.7756, indicating that metadata conditioning improves detection performance on this dataset. The combined score (0.8071) outperforms both the pixel component (0.8016) and the feature component (0.7425), meaning the two error terms are complementary: combining them produces a better ranking than either alone.

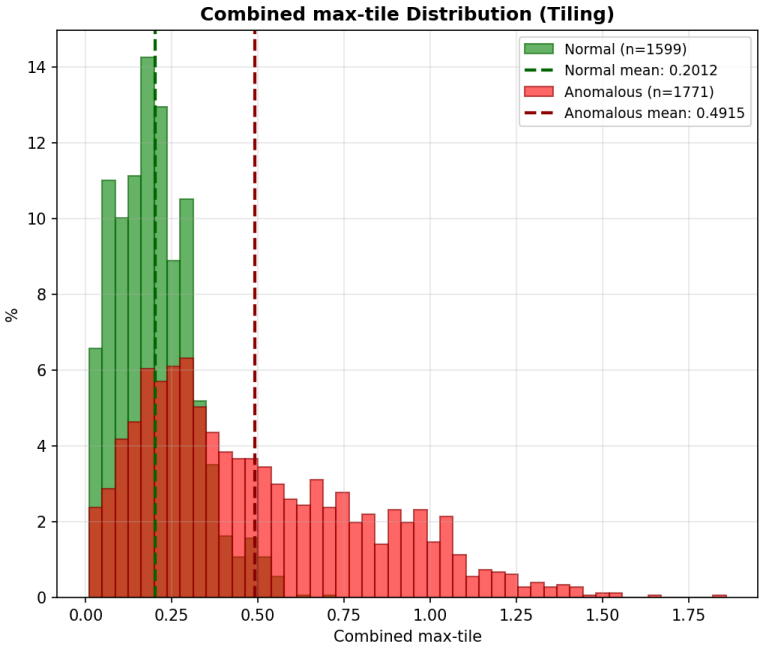


Figure 5.12: Score distribution of the combined max-tile anomaly score for the metadata-conditioned model. The separation between nominal and anomalous distributions is larger than in the baseline (Figure 5.1), reflecting the improved overall AUROC.

The score distribution in Figure 5.12 shows a clearer separation between nominal and anomalous images compared to the baseline (Figure 5.1), which is consistent with the improved AUROC.

Table 5.25 shows performance per anomaly type. Lines are the easiest to detect (AUROC = 0.8859), exceeding the baseline value of 0.725, with the feature component (0.8966) driving most of the discrimination. Noise is detected primarily through the pixel component (AUROC-px = 0.9599), while its feature component (0.6840) is comparatively weak — the opposite pattern to lines, which suggests that the two anomaly types activate different error pathways. Half-swap anomalies remain below random performance (0.4979), showing that tile position metadata does not help the model detect spatial misalignment in the full-disk image. These results suggest that the metadata conditioning does help the model in capturing useful features for anomaly detection, and that the model is able to use these features to capture anomalies better than the baseline.

The per-channel breakdown in Table 5.27 shows strong performance across most channels. For MET09, WV0620 remains the best channel (AUROC-px = 0.8896), closely matching its baseline value of 0.900, while VIS0060 is the weakest (0.7023), which is consistent with the baseline. For MTSAT1, WV0670 remains strong at

Table 5.25: Per-anomaly-type f-AnoGAN detection with metadata conditioning on the MTSAT/MSG test set. Each type is evaluated against all 1,599 nominal images. Best value per column is **bold**.

Anomaly type	AUROC-px-max	AUROC-ft-max	AUROC-cb-max	n
Noise	**0.9599**	0.6840	**0.9271**	319
Patches	0.8712	**0.9102**	0.9018	284
Lines	0.8480	0.8966	0.8859	311
Border black patches	0.8232	0.8264	0.8358	282
Celestial body	0.7534	0.6280	0.7311	319
Half swap	0.5069	0.4921	0.4979	256

Table 5.26: Per-satellite pixel-level anomaly detection with metadata conditioning on the MTSAT/MSG test set.

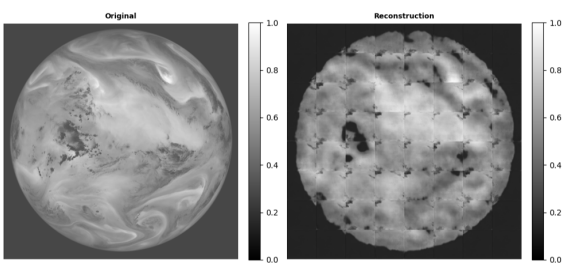
Satellite	AUROC-px-max	AUROC-ft-max	n_{nominal}	n_{anom}
MET09	**0.8051**	0.7232	1100	1207
MTSAT1	0.7951	**0.8057**	499	564

Table 5.27: Per-channel pixel-level anomaly detection with metadata conditioning on the MTSAT/MSG test set. Best AUROC-px per satellite is **bold**; worst is *italic*.

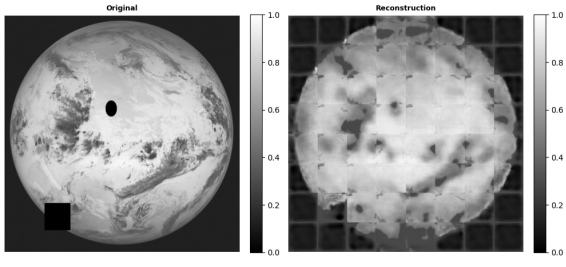
Satellite	Channel	AUROC-px-max	AUROC-ft-max	n_{nominal}	n_{anom}
MET09	WV0620	**0.8896**	0.8031	100	114
MET09	IR1340	0.8712	0.7601	100	116
MET09	IR0970	0.8783	0.7824	100	115
MET09	IR1200	0.8679	0.7066	100	114
MET09	IR1080	0.8424	0.7987	100	113
MET09	IR0870	0.8463	0.7672	100	116
MET09	WV0730	0.8270	0.7202	100	116
MET09	VIS0080	0.7368	**0.8036**	100	99
MET09	IR0160	0.7650	0.8015	100	104
MET09	IR0390	0.7296	0.7333	100	114
MET09	VIS0060	*0.7023*	*0.7547*	100	86
MTSAT1	WV0670	**0.8968**	0.8138	100	114
MTSAT1	IR1200	0.8963	**0.8527**	100	113
MTSAT1	IR1080	0.8469	0.8225	100	114
MTSAT1	VIS0072	0.7181	0.8037	99	108
MTSAT1	IR0375	*0.6721*	*0.7948*	100	115

0.8968, close to its baseline value of 0.896, while IR1080 recovers to 0.8469 from 0.457 — the weakest channel under the baseline. Overall, the metadata-conditioned model achieves more consistent performance across channels than the baseline, with fewer extreme low outliers.

Figure 5.13 shows an example reconstruction from the metadata-conditioned model. The reconstruction is somewhat sharper than the baseline, with residual grid artefacts still visible. On the anomalous image, the block artefact is absent from the reconstruction, confirming that the encoder projects the input onto the learned nominal manifold regardless of the conditioning signal, and that the anomaly score is driven by the resulting reconstruction mismatch.



(a) Normal image. The reconstruction follows the original closely; residual grid artefacts are still present but reduced compared to the baseline.



(b) Anomalous image containing a rectangular block artefact.

Figure 5.13: Reconstruction examples from the metadata-conditioned model on a normal (left) and an anomalous (right) test image. Each panel shows the original (left) and encoder-driven reconstruction $G(E(\mathbf{x}))$ (right).

5.7.4. Randomised metadata ablation

The results above show that conditioning on metadata improves overall AUROC, but they do not directly prove that the model learned to use the metadata during inference. Because the evaluation always provides correct metadata, a model that ignores the conditioning could produce identical scores. To verify that the metadata is actually used during scoring, the evaluation is repeated with shuffled metadata: within each batch, the satellite family index and scalar metadata vector are permuted across images such that no image retains its own metadata, so each image receives real metadata from a different image in the same batch rather than synthetically generated values. If the model learned meaningful conditioning, feeding incorrect metadata should increase the anomaly score for nominal images, causing AUROC to drop. If AUROC remains unchanged, the model effectively ignores the metadata signal.

Table 5.28: Randomised metadata ablation: overall AUROC with correct vs. shuffled metadata on the full test set (1,599 nominal, 1,771 anomalous images).

Metadata	AUROC-combined	AUROC-pixel	AUROC-feature
Correct	0.8071 (0.7911–0.8217)	0.8016 (0.7856–0.8158)	0.7425 (0.7251–0.7585)
Shuffled	0.7871 (0.7710–0.8015)	0.8016 (0.7856–0.8158)	0.7091 (0.6918–0.7256)

The randomised metadata test shows no change in AUROC for the pixel level when metadata is shuffled across images. For the feature error, there is a small drop in AUROC from 0.7425 to 0.7091. This can be explained by the fact that the pixel error is dominated by the image content, and the metadata is only a small portion of the total error. The feature error, however, is determined from the critic's feature activations of both the original and reconstructed image. Shuffling the metadata degrades the reconstruction, causing these activations to diverge, which results in a small drop in AUROC for the feature component. Overall, the randomised metadata ablation suggests that the model does learn to use the metadata during inference, but is limited to the feature level, making the contribution to the final anomaly score relatively small, which may explain why shuffling has a limited effect on overall AUROC.

6

Discussion

This chapter discusses the results of the experiments and their implications for anomaly detection in geostationary satellites. The strengths and weaknesses of the proposed methods are analysed, and potential improvements are identified. The chapter opens with discussing the baseline performance, fine-grained detection, tile overlap, spectral band sensitivity, transfer learning, and metadata conditioning, then Section 6.7 considers the practical implications for the GIAD pipeline, and the chapter closes with a synthesis that positions f-AnoGAN relative to alternative approaches.

6.1. Baseline model performance

In the baseline results, the f-AnoGAN model achieved an AUROC of 0.7756, which exceeds a random classifier, validating that the method works, while leaving room for improvement. The combined score outperforms both the pixel-only and feature-only scores.

Breaking down the performance per anomaly type it shows that the half swap anomalies cannot be detected successfully. This was partially solved by using overlapping tiling approach, discussed in Section 6.3.

Another anomaly that was hard to detect is the celestial body anomaly. This might be caused by the fact that the celestial body is a smaller anomaly. However, another possible explanation can be found when looking at the performance per spectral band. In that overview it can be seen that the visible light bands, and near infrared bands, have a lower performance compared to the other bands.

A possible reason for the lower performance of the near infrared and visible light bands, compared to the higher wavelength spectral bands, is that they have a day/night cycle, which makes the nominal data more variable. This can make it harder for the model to properly learn to reconstruct this day/night cycle.

This hypothesis is supported by the fact that the water vapour bands have the best performance, and these bands are not influenced by the day-night cycle, but also have less complicated cloud patterns than the infrared and visible light bands.

The performance between the two different satellites present in the dataset is similar, and the MET09 performs slightly better than the MTSAT-1. However, this difference is insignificant, and can possibly be explained by the fact that the MET09 has more training data than the MTSAT-1, which can mean that the model is better tuned to the MET09 satellite, and therefore performs better on it.

When comparing the autoencoder baseline to the f-AnoGAN approach, the f-AnoGAN achieves a higher overall AUROC (0.77 versus 0.71). Despite the improvement not being substantial, the f-AnoGAN approach demonstrates a more robust detection capability. When looking into the per-anomaly and per-channel performance, f-AnoGAN performs more consistently across different anomaly types and channels, while the autoencoder has some categories it detects very well and others it detects poorly. This shows that f-AnoGAN generalises better across the evaluation conditions, while the autoencoder is more sensitive to specific anomaly types and channels.

Across all experiments, it became evident that in general, the WGAN was able to find and learn how to recreate the nominal tiles in the dataset reliably. The encoder on the other hand had a hard time to reverse the generator, and find the right latent space encoding for the images. This led to weak reconstructions, only representing high-level textures and shapes. The root cause is likely that the encoder is trained on a stale generator — one that was never optimised with easy invertibility in mind. Because of this, the encoder

can have a hard time reversing the generator's latent space. A possible solution for this is to implement the GANomaly architecture [3].

6.2. Fine-grained anomaly detection

As can be expected, the fine-grained anomaly detection experiments have a lower performance than the baseline experiment, where the absolute max tile score achieves a combined AUROC of 0.6053. This can be visually explained by observing the reconstructions. Figure 5.6 shows an example of the reconstruction and pixel error map from the baseline model on a fine-grained anomalous image. The anomaly is a small patch of black pixels, which produces only a localised spike in the pixel error map. However, when looking at the whole error map, due to the overall smooth reconstruction, there are more pixel error spikes in areas without anomalies, making it hard to distinguish between nominal and anomalous images. The reason absolute max tile does perform a bit better probably has to do with the fact that fine-grained anomalies do not raise the mean error of a tile that much, but the spike in absolute error in the raw error maps is more likely to be the highest error tile, giving a better performance than the normalised max tile score.

6.3. Effect of tile overlap

When training the f-AnoGAN model on a dataset with tiles with overlap, there is a small improvement in performance. The combined AUROC improved from 0.7756 (baseline) to 0.8155 for the overlap-trained model. When looking at the reconstructions of the image (Figure 5.8), the grid artefact mostly disappeared compared to the baseline model. This is also reflected in the overall error of both the nominal and anomalous images being lower compared to the baseline, while the difference between the nominal and anomalous mean slightly increases, giving better separation. The standard deviation of the nominal score distribution also decreases, making the nominal and anomalous images even better separable. Notably, the half-swap anomaly improved from a combined max-tile AUROC of 0.495 (baseline, Table 5.4) to 0.714 (overlap model, Table 5.15), the largest per-type gain across all experiments. This suggests that extra global context, due to more tiles, and shared data between tiles, enables the model to detect misalignment anomalies. One thing to note here is that this improvement may only occur when the misalignment falls within overlapping tile regions; further experiments would be needed to confirm this.

The controlled comparison (Section 5.3.4) shows that applying overlap at evaluation time to the original baseline model does not reproduce this gain — the combined AUROC dropped from 0.7756 to 0.6759. This confirms that the improvement comes from the model learning to generate smoother tile transitions during training, not from the blending step at evaluation time.

6.4. Spectral band sensitivity

The results of the VIS and NIR experiments show that a good dataset is important for the performance of f-AnoGAN. The model suffered from mode collapse, caused by the lack of sufficiently variable training data, showing that the model is not able to learn a good representation of only NIR and VIS bands. f-AnoGAN therefore proved unsuitable for anomaly detection in spectral bands affected by the day-night cycle under this setup. A possible solution might be to investigate better, more balanced datasets for these bands, where a smaller portion consists of uniform background tiles. The metadata experiment showed improvements in the performance of the NIR and VIS bands, where especially the feature max score improved to scores ranging from 0.63 to 0.70 in the baseline to 0.75 to 0.80 in the metadata experiment. This shows that the metadata conditioning is able to help the model to learn better representations of the nominal data.

When removing the NIR and VIS channels from the dataset and training the model on only the TIR and WV bands, the model achieved a combined AUROC of 0.7962, comparable to the full-band baseline (0.7756). The per-anomaly-type pattern is consistent with the baseline: lines and noise are the easiest to detect, half-swap and celestial body remain near-random. Water vapour channels again show the strongest per-channel performance, while the long-wave infrared channels (IR0390, IR0375) are the weakest. Restricting the dataset to TIR and WV bands therefore does not substantially change the overall detection capability, which suggests that the NIR and VIS bands contributed little to the baseline AUROC despite their presence in training.

6.5. Transfer learning

The transfer learning experiment shows that generalising to new unseen satellites is less trivial than anticipated. The transfer learning results show that a pretrained model, that is retrained with 100 images per chan-

nel of the new target domain, performs worse than a separate model trained on the target domain images only. Visual inspection of the generator outputs and reconstructions (Figure 5.9) confirms that the generator successfully learns the new image structures, but that the encoder fails to produce meaningful latent representations for the new satellites.

A possible explanation lies in the two-stage training structure of f-AnoGAN. The encoder is trained to invert one specific frozen generator, so it does not learn a general inversion skill, it learns to navigate that generator's latent space. When the generator is fine-tuned on GOES-11, its mapping changes, and the encoder's weights are immediately out of date. Re-training the encoder on GOES-11 data alone does not fix this: the generator's latent space was based off a much larger MTSAT/MSG dataset, so the encoder only ever sees a small part of it. A stand-alone encoder trained from scratch does not have this problem, since it only needs to cover the GOES-11 portion of the latent space, which is exactly what its training data provides. A practical fix would be to re-train the encoder on a mix of pre-training and target-domain images, exposing it to the full latent space rather than just the target-domain corner this approach was considered out of scope for this thesis and left as future work.

6.6. Metadata conditioning

The metadata-conditioned model achieved a combined AUROC of 0.8071, compared to 0.7756 for the baseline, this improvement is just outside the confidence intervals, suggesting that the metadata conditioning contributed slightly to the performance increase.

Despite the overall improvement, the half-swap anomaly remained below random performance (AUROC = 0.5069), showing that conditioning on tile position did not help the model detect spatial misalignment. This is somewhat surprising, because tile position metadata should in theory provide the generator with a spatial prior: when given a centre-tile position, the generator is encouraged to produce images that look like centre tiles. An anomalous tile appearing at the wrong position would then mismatch this prior and produce a higher reconstruction error. In practice, however, this mechanism appears too weak to overcome the reconstruction error from nominal pixel patterns.

This is supported by the randomised metadata ablation (Table 5.28): shuffling metadata across images left the pixel AUROC unchanged at 0.8016, while the feature AUROC dropped slightly from 0.7425 to 0.7091, indicating that the model uses metadata only weakly and exclusively through the critic's feature activations, and that the overall improvement over the baseline is therefore not attributable to the metadata conditioning itself, but rather to differences in training dynamics under the metadata experiment setup. Stronger conditioning mechanisms, such as FiLM [39], which inject metadata at every layer rather than only at the latent vector, may provide sufficient weight to make the conditioning effective.

6.7. Practical implications

Comparing the GIAD method to the new f-AnoGAN approach is not straightforward, because of the structural differences between the two methods. The GIAD approach is a deterministic rule-based approach, where the anomalies need to be known, and a detector is created for each anomaly individually. f-AnoGAN is an unsupervised anomaly detector, that is able to find anomalies in datasets, without prior knowledge about the anomalies. The f-AnoGAN gives a continuous score to an image based on how anomalous it looks, and then based on a threshold, the image can be classified as anomalous. Due to the unsupervised learning approach, the f-AnoGAN does not need information about anomalies during training, enabling it to detect known and unknown anomalies. The GIAD approach, as said, is deterministic: an image is anomalous or not, and is not able to detect unknown anomalies. Putting the strengths and weaknesses of both detectors next to each other, it shows that these methods can be used complementarily. F-AnoGAN, being unsupervised and able to generalise to new datasets, can work as a first pass to help the human expert detect anomalies in new datasets, after which GIAD is able to classify the anomalies. When f-AnoGAN has identified anomalous images, together with the anomaly locations, a human expert can check the output to see if new anomalies occurred, and whether the current GIAD rules need additions or updates. Once this is done, GIAD can be executed on the whole dataset to capture all anomalies with higher certainty.

Looking at the impact of the results on the practical usability of the detection method. With an AUROC of 0.7756 and an F1 of 0.7256 the baseline method shows promise as a part of the anomaly detection pipeline, but it is not good enough to function as a stand-alone detector in a practical application. Given that the overlap and metadata experiments showed improvements on this score to 0.8155 and 0.8071 respectively, it is possible that with the combined improvements the method becomes more reliable in the future, however,

since these scores are based on synthetic data, it is yet uncertain how the method will perform on real data. Lastly, since the f-AnoGAN, despite some level of anomaly localisation is still a black-box system, it is advised to keep a human in the loop to check anomalies, and implement more robust and deterministic anomaly detection methods, like the rule-based approach taken in GIAD.

When discussing the practical applicability of the method, it is important to look at the potential users of the system. The main groups that are impacted by the anomaly detection system are the satellite operators, who are responsible for maintaining the satellites and ensuring data quality; the end-users, like climate researchers, who rely on the data for their research; and lastly the maintainers of the anomaly detection pipeline, who are responsible for keeping the system working. These three groups all have different needs and requirements for the anomaly detection system. For the satellite operators, a system with a low false positive rate is important, since the time spent by domain experts validating anomaly alerts should be minimised — with a baseline F1 of 0.7256, a conservatively set threshold is necessary to avoid overwhelming reviewers with false alarms. For the end-users however, it is more important that all anomalies are detected, since a false negative can silently corrupt research results; this is particularly relevant given that spatial misalignment anomalies such as the half-swap remain near-random (AUROC ≈ 0.50), meaning the method should be combined with a deterministic second pass for archival use. Lastly for the maintainers of the detection pipeline, reliability and computational efficiency are the primary concerns: since the pre-trained model does not transfer to unseen satellites, each new instrument requires a full re-training run, which should be factored into operational planning.

Because these three groups have conflicting requirements — high precision for operators, high recall for end-users — no single threshold configuration can serve all of them simultaneously. Deployment decisions should therefore include explicit threshold calibration for each operational context.

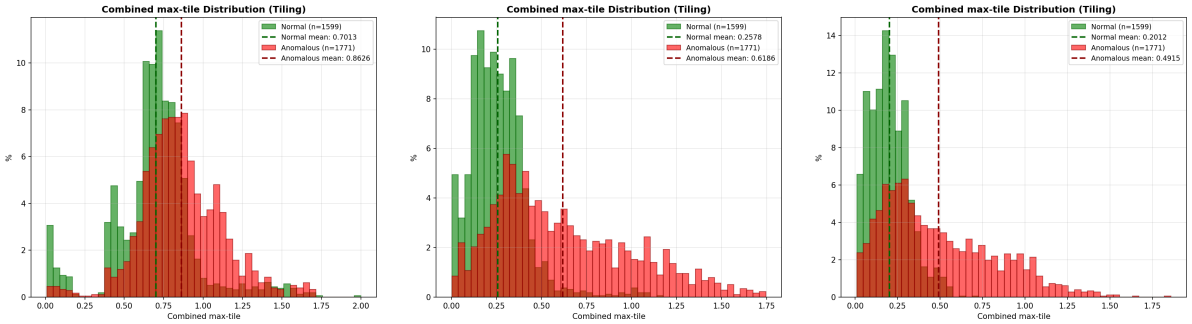
Lastly, it should be mentioned that the anomalies used in this research are mostly synthetic. This means that the performance of the detectors on real-world data might not be the same, because real anomalies differ from the synthetic ones. The synthetic anomalies were created with the goal of evaluating overall performance across different kinds of anomalous pixels and regions, and to test whether hypothesised weaknesses could be confirmed by creating anomalies that support those hypotheses.

6.8. Synthesis

The experiments showed that the f-AnoGAN method showed some promise for anomaly detection in geostationary satellite images, but also showed some weaknesses. The method is able to detect large anomalies and structural anomalies like the broken lines or noise anomalies. However, there are weaknesses: the base model suffers from the inability to detect misalignment anomalies, like the half-swap. It is worth noting that the half-swap is a synthetic worst-case; real geometric misalignments in the dataset (Chapter 2) are typically less severe, so the practical impact of this weakness is limited. When looking at the reconstructions, it can also be seen that only the high-level features are reconstructed well. The model therefore cannot reliably detect fine-grained anomalies. The tiling approach, while giving room for the evaluation of higher resolution images, comes at the cost of losing global context, even with added metadata context. As mentioned, improving the way the metadata is added might give it more weight and solve this problem. Another possible fix is training a model on downsized non-tiled images, which does retain global context and would be able to recognise large misalignment anomalies covering the complete image. Adding this as a first pass to the anomaly detection pipeline would be a remedy to this weakness. The transfer learning experiment showed that the model is not able to generalise well to new unseen satellites, due to architectural limitations of the f-AnoGAN. This means that for new satellites, new models need to be trained, which makes it harder to quickly be able to detect anomalies in new satellites.

The base performance of the f-AnoGAN can be improved by introducing metadata, which gives the model more information about the input tiles, and thus enabling the model to learn a better representation of the nominal data. Additionally, adding overlapping sections between the tiles also improves the model by giving it more spatial context, which lowers the reconstruction error of the image as a whole when tiles are being put together. This then results in lower error scores on nominal images, while the error on anomalous images stays roughly the same. Because of this, the distributions become more separate, enabling better detection, as shown in Figure 6.1.

Lastly, it is important to reflect on how the discussed f-AnoGAN method is positioned relative to other anomaly detection methods. The baseline AE implementation concluded that AE reconstructions were too smooth to preserve the anomaly signal in high-resolution multi-band imagery, this is partially solved by f-



(a) Baseline (b) Tile overlap (c) Metadata

Figure 6.1: Combined max-tile score distributions for nominal and anomalous images across the three main model variants. Both the overlap and metadata models show a tighter nominal distribution and greater separation between the two classes compared to the baseline.

AnoGAN, that uses more complex scoring methods based of the critics latent space. At the same time, the transfer learning findings suggest a structural weakness of the two-stage training setup: because the encoder is trained to invert one specific generator, f-AnoGAN is harder to adapt to new domains than a single-stage model such as GANomaly [3] or an AE, both of which couple generation and encoding in a single pass. Diffusion-based detectors [28] would likely produce higher-quality reconstructions, and probably sharper anomaly maps. However, this comes with an inference cost that scales with the number of denoising steps, making them less applicable to real-time screening of large archives. Considering this, the f-AnoGAN is located on the middle ground for structural, image-level anomaly detection in multi-domain satellite data, but that single-stage architectures should be prioritised if domain transfer or fine-grained localisation are primary requirements.

7

Conclusion

In this chapter, the main findings of the thesis, together with the answers to the research questions, are presented. The recommendations based on the findings are given, and future research directions are outlined to further improve anomaly detection in geostationary satellite data.

7.1. Answer to Research Questions

First, it is good to look back at the research questions posed at the start of the thesis, and see how these questions are answered by this research. The subquestions will be discussed first, and then the primary research question will be answered based on the answers to the subquestions.

1. What anomaly detection algorithms are able to detect anomalies of different nature (smaller anomalies like hot pixels and bigger anomalies like corrupt lines) in geostationary satellite images?

From the baseline experiments in Section 5.1.3 and the fine-grained anomaly detection experiments in Section 5.2, it can be concluded that the f-AnoGAN is able to detect structural anomalies like missing earth parts, broken scan lines or noisy parts in the image. However, the performance on fine-grained anomalies is close to random performance. This means that the f-AnoGAN is not a reliable anomaly detector for use cases where it is important to detect fine-grained anomalies, like hot pixels. One advantage of the f-AnoGAN is that it is able to detect novel anomalies that are unknown to the model and user. The results also showed that f-AnoGAN augmented with metadata, and trained on overlapping tiles for more global context, achieved improved performance, where the metadata-augmented model had an AUROC of 0.8071 and the overlapping tiles an AUROC of 0.8155.

2. How can anomaly detection across generations of GOES, Himawari and Meteosat satellites be handled?

The baseline, metadata and overlap experiments showed that the f-AnoGAN is able to learn to detect anomalies when it is trained on a dataset containing multiple satellites and channels. However, looking into domain adaptation, as done in the transfer learning experiment in Section 5.6, it is shown that this does not work straightforwardly. Zero-shot transfer to GOES-11 achieved a combined AUROC of 0.618, and fine-tuning on 100 images per channel improved this to 0.687. However, a model trained from scratch on the same volume of GOES-11 data achieved 0.696, showing that the pre-trained MSG/MTSAT manifold does not provide a beneficial initialisation for an unseen satellite. Noise and line anomalies were the most transferable anomaly types; half-swap and celestial body detection remained near-random across all fine-tuning conditions. So while the f-AnoGAN is able to generalise across different satellites and spectral bands, when these are all present in the training data, it does not support transfer learning to new satellites. This means that for new satellites, new separate models need to be trained.

3. Where does a learned anomaly detector complement the rule-based GIAD approach, and how can the two methods support each other in an anomaly detection pipeline?

As discussed, the GIAD approach and f-AnoGAN approach have different strengths and weaknesses. F-AnoGAN is able to detect structural anomalies such as noise patterns, broken scan lines, and large patches, and can flag novel anomaly types that are never encountered before, and thus not yet covered by GIAD rules, since it does not require prior knowledge of what an anomaly looks like. GIAD, on the other hand, reliably catches the errors for which a detector has been developed, including fine-grained pixel faults such as hot pixels and spatial misalignment anomalies, both of which f-AnoGAN was unable to reliably detect. By com-

bining the two approaches, it is possible to create a more automated anomaly detection pipeline, reducing human labour. Where before, human experts had to go through the data manually to find anomalies, now the f-AnoGAN model can be utilised to make a first selection of anomalous-looking images. These images can then be inspected to see if new anomalies occur and whether new rule-based detectors need to be created. The GIAD method could then be used to find anomalies the f-AnoGAN might have missed and give high confidence in the found anomalies.

Primary Research Question:

How can anomaly detection methods be designed to detect both pixel- and image-level anomalies in different geostationary satellite imaging systems without requiring labelled anomaly examples? The f-AnoGAN framework, trained on nominal data alone, successfully detects structural anomalies like broken scan lines, noise and black patches across multiple satellites and spectral bands without requiring labelled anomaly examples. Reconstruction error alone proved insufficient to reliably detect anomalies, given the high resolution and variability of the training data; combining pixel-space and discriminator-feature-space errors improved detection breadth. Incorporating overlapping tiles and satellite metadata each improved performance further (AUROC 0.8155 and 0.8071 respectively). The method falls short for fine-grained pixel-level faults (AUROC = 0.6053) and spatial misalignment anomalies such as the half-swap (AUROC $\approx$ 0.50), neither of which was resolved by the overlap or metadata experiments, and both of which require either higher-resolution context or stronger conditioning mechanisms than the baseline tiling approach provides.

7.2. Recommendations

Based on the answers provided to the research questions, some recommendations on how to use and interpret this research will be provided. The method can be used as a first-pass detector to flag larger, novel anomalies in unseen datasets, from which new rule-based detectors can be derived, but it should not be relied upon as a stand-alone solution. The first recommendation is to evaluate the trained models on a test set of real anomalies, to get a better understanding of the real world performance of the model. This will show whether the model has any real world applicability, and will also show whether the model is able to detect real anomalies, or that it is only able to detect the synthetic anomalies. To fully evaluate the real world applicability of the methods, it is also important to consult with domain experts, to see if the found anomalies and the amount of false positives, are acceptable for use in practice. This will also give insights into how the model can be improved, and which anomaly types are the highest priority to detect. Then, this thesis describes different methods to improve the performance of the f-AnoGAN method, such as adding metadata, and using overlapping tiles. These methods were tested separately, but it would be interesting to see how the model performs when both methods are combined. This could potentially lead to even better performance, as the model would have more information about the global context of the image, and would also have more information about the satellite and channel datasets, to help find new anomalies.

7.3. Limitations

The metadata experiment showed that providing metadata to the input images improved the performance of the model. However, the randomised metadata test showed that during inference, the metadata had no influence on the anomaly scores. To address this, techniques like FiLM [39] could be tested. With stronger metadata conditioning it would also be interesting to see if the model is able to identify errors in metadata labels, for example a WV image that is labelled as VIS.

The f-AnoGAN showed improvement when overlapping tiles were implemented, and when metadata was added to the input data. However, these improvements have only been tested separately. In the future, it might be interesting to see how the model improves when both metadata and overlapping tiles are combined into one model.

One pattern that occurred throughout the experiments, which was briefly touched upon in the channel-specific experiments, is the large difference in performance between VIS and NIR channels and TIR and WV bands. The experiments focusing on this difference showed that training exclusively on the VIS and NIR channels introduced the problem of mode collapse, making it impossible to create a good model for these channels alone. It showed that a model trained without the VIS and NIR channels achieved slightly better results. Investigating how the f-AnoGAN method could better learn images that are influenced by the day-night cycle — for example through more targeted metadata conditioning or by increasing their presence in the training data — might be interesting to further improve the overall performance of the model.

The transfer-learning experiment showed that the pre-trained model with MET09 and MTSAT1 data did

not generalise well to the unseen GOES-11 data, and that training a new model from scratch on the same data volume, outperformed transfer learning. An identified cause for this is the retraining needed for the encoder in f-AnoGAN. A test that might be interesting to do in the future, is to test transfer learning, where the WGAN is retrained on a small dataset of only unseen images, and the encoder then retrained on a dataset containing a subset of the original training data, in addition to some unseen data. This might solve the problems encountered now in the encoder training.

In this thesis, the model evaluations were done with synthetic anomalies, which were inspired by real anomalies, but were not real anomalous images. This means that the performance shown in this thesis might give a skewed impression of the real performance of the model on real anomalies. The synthetic anomalies, while designed to represent real anomalies, were also designed to test different detection capabilities of the model. This means that some anomalies might be overrepresented in the test set, or that some anomalies are spanning a bigger part of the image than the real anomalies. To limit this, the fine-grained anomaly detection experiment was designed to also get a better understanding of the performance of the model on smaller anomalies. However, this does not mean that these fine-grained anomalies are perfect representations of real anomalies. To really get the real world performance of the tested models, evaluation on a test set containing real anomalies is a missing validation. Doing this validation will give a better understanding of the real performance of the model, and will also show whether the model is able to detect real anomalies, or that it is only able to detect the synthetic anomalies. For the implementation of the model in an anomaly detection pipeline, this evaluation on a realistic test set is important for other reasons as well. When calculating the anomaly scores, the detection threshold, and the score calibration are based on the test set. So if the test set is not representative of the real world, the threshold and score calibration might not be optimal for real world use, which could lead to worse performance in practice.

The dataset used in this research, while consisting of two distinct satellite programmes, and 14 distinct channels, is still relatively small, with only 10k images and a total of 44 channels. The results shown in this paper while showing that the methods are able to generalise over a diverse dataset, might turn out different when a bigger even more diverse dataset is used. Also, this means that if the method is to be used in practice, and for the complete range of satellites and channels, a new model would have to be trained on a bigger dataset containing data from all satellites and channels. This would require more computational resources, and more time to properly train the model. However, it would also likely lead to better performance, as the model would be able to learn from more data, and would be able to learn a better data representation.

The overlapping experiment showed that adding more global context to the training set improved the model performance. However, the model still only receives context about one tile, during evaluation, and has no information about the temporal context of the image. Since the satellites capture images at regular intervals, and cloud patterns change smoothly over time, it might be interesting to see if the model performance can be improved by giving it more temporal context. This could be done by giving the model a sequence of images as input, instead of just one image, or by giving the model a sequence of tiles as input, instead of just one tile. This would give the model more information about the global context of the image, and might help it to better detect anomalies that are not visible in a single tile, but are visible in a sequence of tiles.

The research is also limited by the available hardware. The runtime of the experiments was already quite long for the base model, and the model size, image size, and batch size were all constrained by the available GPU memory. As with many deep learning methods, performance increases with more training data and larger models. In the case of high-resolution anomaly detection, more memory would enable larger tile sizes and the use of less aggressively downsampled images.

7.4. Future Work

The limitations described above also give rise to some interesting future research directions. Some of these are directly motivated by the results of this research, while others are more broadly motivated by the field of anomaly detection in satellite images.

For the metadata conditioning, determining what kind of metadata increases model performance the most should be investigated. In this research, the added metadata included the satellite, channel tile location and time of acquisition. This metadata could be augmented by for example adding satellite location, or time of day. Instead of adding more metadata, it could also be interesting to see if the model performance can be improved by giving more general metadata, for example only indicating the channel type (VIS, IR, WV) instead of the specific channel, or a approximate time of day (day, night, twilight) instead of the exact time of acquisition. This might make it easier for the model to learn useful relations over the metadata, and might

also make the model more robust to errors in the metadata labels. Implementing stronger conditioning techniques such as FiLM [39] or CC-CLIP [35], to see if this can make the model more sensitive to the metadata, and whether this can improve the performance of the model is a promising next step. The FiLM method works by applying a learned transformation to the feature maps of the models. By applying this satellite metadata (Satellite ID, channel, tile location) to all the intermediate feature maps into the generator, encoder and critic of the f-AnoGAN, the latent representation could be made aware of the imaging context. The CC-Clip method works by comparing a subject to the context it is in, for example a runner on a track, or a runner on a highway. In both cases the subject is a runner, which is not anomalous. The runner only becomes anomalous in the incorrect context, like the highway. This idea can be applied to the satellite images as well, where the tile is the subject, and the metadata is the context. Then an image of a watervapor tile, with metadata of a visual light channel is anomalous. By using this method, the model might be able to better learn the relationship between the metadata and the image, and this might lead to better performance.

This thesis showed that the current f-AnoGAN method does not support transfer learning to new satellites. Future research could investigate how to make the method more suitable for transfer learning, for example by retaining the WGAN on a small dataset of only unseen images, and then retraining the encoder on a dataset containing a subset of the original training data, in addition to some unseen data. This might solve the problems encountered now in the encoder training. Changing the architecture of the model, to remove the need for two training stages is another good approach. An example of such an approach is GANomaly [3]. Here, instead of training a WGAN, and then a separate encoder, the model is trained in one stage, this done by training a normal GAN, with an autoencoder as the generator, which enables the generator to accept images as an input, removing the need to later learn the image-to-latent mapping in a separate encoder training stage. This might make the model more suitable for transfer learning.

Looking beyond the specific method used in this research, there are also some interesting future research directions that could be explored. One of these is the use of time series analysis for anomaly detection.

In this thesis, and in previous work on anomaly detection in satellite images, the main focus has been on per-image anomaly detection. Another interesting approach could be based on time series analysis. Since satellites capture images at regular intervals, and cloud patterns change smoothly over time, anomalies might be detected by identifying large deviations in this temporal data. Adding temporal context to the model, by for example combining the f-AnoGAN method with a recurrent neural network or a LSTM [19, 62, 54], which are designed to handle sequential data, and they can learn to identify patterns and anomalies in time series data, by being able to process and embed information of sequential data. By combining the f-AnoGAN method with a recurrent neural network(RNN) or a LSTM, the model could learn to identify anomalies based on pixel or region changes over time, in addition to the image context. This could potentially lead to better performance, as the model would possibly be able to detect anomalies, because of sudden extreme changes in pixel or region values over time. Adding the temporal context of the images would mean that input images have to be ordered sequentially, which would require a change in the data loading and training process.

With the high resolution of the available data, hierarchical anomaly detection could enhance the detection of anomalies by providing a multi-scale approach. In this approach, a coarse full-image model first screens for global anomalies such as half-swaps, misaligned scans and big structural anomalies, where the current tile-level model lacks the global context to reliably detect them. Then smaller high resolution tiles are processed for fine-grained anomaly detection, such as hot pixels and small noisy regions, where the high resolution images give more precise information about the local context of the image. A two-stage hierarchical approach, as explored in high-resolution industrial inspection [45], could address this gap without requiring a fundamentally different architecture.

Beyond the technical scope of this research, the applicability of automated anomaly detection in satellite imagery should also be evaluated. Geostationary satellites like Meteosat have been operational since the 1970s, producing continuous archives that are increasingly used for climate research and reanalysis. Ensuring the data quality of these archives, by flagging corrupted or anomalous images, is important for the reliability of climate products that are derived from these archives. Manual inspection of these archives by domain experts is time-consuming and is not scalable, taking into account the volume of data and the increasing number of satellite programmes. A method like f-AnoGAN, integrated into an operational pipeline at organisations such as EUMETSAT or S[&]T, could reduce the manual work, and enable extra quality control across satellite programmes. As the field of earth observation continues to grow, with more satellites, generating more and higher resolution data, the need for automated anomaly detection will only increase.

Bibliography

[1] *3. The Geostationary Orbit*. https://www.esa.int/Education/3._The_geostationary_orbit. (Visited on 04/20/2026).

[2] *ABI | GOES-R Series*. https://www.goes-r.gov/spacesegment/abi.html. (Visited on 05/13/2026).

[3] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. *GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training*. Nov. 2018. DOI: 10.48550/arXiv.1805.06725. arXiv: 1805.06725 [cs]. (Visited on 12/04/2025).

[4] Jinwon An and Sungzoon Cho. "Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability". In: 2015. (Visited on 11/17/2025).

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. https://arxiv.org/abs/1701.07875v3. Jan. 2017. (Visited on 01/08/2026).

[6] Paul Bergmann et al. "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 9584–9592. ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00982. (Visited on 02/10/2026).

[7] Antonia Creswell et al. "Generative Adversarial Networks: An Overview". In: *IEEE Signal Processing Magazine* 35.1 (Jan. 2018), pp. 53–65. ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2765202. (Visited on 11/17/2025).

[8] Florinel-Alin Croitoru et al. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023), pp. 10850–10869. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3261988. (Visited on 11/18/2025).

[9] Thomas Defard et al. *PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization*. Nov. 2020. DOI: 10.48550/arXiv.2011.08785. arXiv: 2011.08785 [cs]. (Visited on 01/28/2026).

[10] Sarah M. Erfani et al. "High-Dimensional and Large-Scale Anomaly Detection Using a Linear One-Class SVM with Deep Learning". In: *Pattern Recognition* 58 (Oct. 2016), pp. 121–134. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2016.03.028. (Visited on 06/05/2026).

[11] Rebekah Esmaili et al. *Anomaly Detection in Satellite Imagery Using Convolutional Neural Networks*. Mar. 2025. DOI: 10.13140/RG.2.2.10609.19046.

[12] M. R. Gauthama Raman, Wenjie Dong, and Aditya Mathur. "Deep Autoencoders as Anomaly Detectors: Method and Case Study in a Distributed Water Treatment Plant". In: *Computers & Security* 99 (Dec. 2020), p. 102055. ISSN: 0167-4048. DOI: 10.1016/j.cose.2020.102055. (Visited on 11/17/2025).

[13] *GOES East (GOES 16) | National Oceanic and Atmospheric Administration*. https://www.noaa.gov/jetstream/goes_east. (Visited on 05/13/2026).

[14] Dong Gong et al. "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1705–1714. (Visited on 11/17/2025).

[15] Ian J. Goodfellow et al. *Generative Adversarial Networks*. June 2014. DOI: 10.48550/arXiv.1406.2661. arXiv: 1406.2661 [stat]. (Visited on 01/27/2026).

[16] M.A. Hearst et al. "Support Vector Machines". In: *IEEE Intelligent Systems and their Applications* 13.4 (July 1998), pp. 18–28. ISSN: 2374-9423. DOI: 10.1109/5254.708428. (Visited on 11/18/2025).

[17] Frank Hogervorst. "D2.7 Database Design Document". In: (June 2025).

[18] Frank Hogervorst and Jacob Senior-Williams. "Algorithm Theoretical Basis Document". In: (June 2025).

[19] Stephanie Hyland, Cristóbal Esteban, and Gunnar Rätsch. "Real-Valued (Medical) Time Series Generation with Recurrent Conditional GANs". In: (Feb. 2018). (Visited on 05/19/2026).

[20] Seungbin Ji et al. “Multi-View Masked Autoencoder for General Image Representation”. In: *Applied Sciences* 13.22 (Nov. 2023). ISSN: 2076-3417. DOI: 10.3390/app132212413. (Visited on 12/03/2025).

[21] Watson Jia, Raj Mani Shukla, and Shamik Sengupta. “Anomaly Detection Using Supervised Learning and Multiple Statistical Methods”. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. Dec. 2019, pp. 1291–1297. DOI: 10.1109/ICMLA.2019.00211. (Visited on 05/01/2026).

[22] Shehroz S. Khan and Michael G. Madden. “One-Class Classification: Taxonomy of Study and Review of Techniques”. In: *The Knowledge Engineering Review* 29.3 (June 2014), pp. 345–374. ISSN: 0269-8889, 1469-8005. DOI: 10.1017/S026988891300043X. (Visited on 11/18/2025).

[23] Konstantin V. Khlopenkov and David R. Doelling. “Development of Image Processing Method to Detect Noise in Geostationary Imagery”. In: *Image and Signal Processing for Remote Sensing XXII*. Vol. 10004. SPIE, Oct. 2016, pp. 599–607. DOI: 10.1117/12.2241544. (Visited on 11/14/2025).

[24] Christina Köpken. “Solar Stray Light Effects in Meteosat Radiances Observed and Quantified Using Operational Data Monitoring at ECMWF”. In: *Journal of Applied Meteorology and Climatology* 43.1 (Jan. 2004), pp. 28–37. ISSN: 1520-0450, 0894-8763. DOI: 10.1175/1520-0450(2004)043<0028:SSLEIM>2.0.CO;2. (Visited on 09/30/2022).

[25] Haoyuan Li and Yifan Li. “Anomaly Detection Methods Based on GAN: A Survey”. In: *Applied Intelligence* 53.7 (Apr. 2023), pp. 8209–8231. ISSN: 1573-7497. DOI: 10.1007/s10489-022-03905-6. (Visited on 11/17/2025).

[26] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest”. In: *2008 Eighth IEEE International Conference on Data Mining*. Dec. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17. (Visited on 12/10/2025).

[27] Jing Liu et al. *A Survey on Diffusion Models for Anomaly Detection*. Feb. 2025. DOI: 10.48550/arXiv.2501.11430. arXiv: 2501.11430 [cs]. (Visited on 11/18/2025).

[28] Victor Livernoche et al. *On Diffusion Modeling for Anomaly Detection*. Mar. 2025. DOI: 10.48550/arXiv.2305.18593. arXiv: 2305.18593 [cs]. (Visited on 05/01/2026).

[29] Runzhou Mao et al. *Sequential PatchCore: Anomaly Detection for Surface Inspection Using Synthetic Impurities*. Jan. 2025. DOI: 10.48550/arXiv.2501.09579. arXiv: 2501.09579 [cs]. (Visited on 01/28/2026).

[30] Stefania Matteoli, Marco Diani, and Giovanni Corsini. “A Tutorial Overview of Anomaly Detection in Hyperspectral Images”. In: *IEEE Aerospace and Electronic Systems Magazine* 25.7 (July 2010), pp. 5–28. ISSN: 1557-959X. DOI: 10.1109/MAES.2010.5546306. (Visited on 11/18/2025).

[31] Diogo Medeiros. *Icon Generation with Conditional Generative Adversarial Networks*. Jan. 2023. DOI: 10.13140/RG.2.2.26065.20320.

[32] W. Paul Menzel and James F. W. Purdom. “Introducing GOES-I: The First of a New Generation of Geostationary Operational Environmental Satellites”. In: *Bulletin of the American Meteorological Society* 75.5 (May 1994), pp. 757–782. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(1994)075<0757:IGITFO>2.0.CO;2. (Visited on 05/21/2026).

[33] *Meteorological Satellite Center (MSC) | Himawari-8/9 Imager (AHI)*. https://www.data.jma.go.jp/mscweb/en/himawar (Visited on 05/13/2026).

[34] *Meteosat Third Generation Instruments | EUMETSAT*. http://www.eumetsat.int/meteosat-third-generation-instruments. Wed, 16/11/2022 - 14:40. (Visited on 05/13/2026).

[35] Shashank Mishra, Didier Stricker, and Jason Rambach. *Conditional Compatibility Learning for Context-Dependent Anomaly Detection*. 2026. DOI: 10.48550/arXiv.2601.22868. arXiv: 2601.22868 [cs.CV]. (Visited on 05/19/2026).

[36] Takeru Miyato and Masanori Koyama. *cGANs with Projection Discriminator*. Aug. 2018. DOI: 10.48550/arXiv.1802.05637. arXiv: 1802.05637 [cs]. (Visited on 03/25/2026).

[37] Asif Ahmed Neloy and Maxime Turgeon. “A Comprehensive Study of Auto-Encoders for Anomaly Detection: Efficiency and Trade-Offs”. In: *Machine Learning with Applications* 17 (Sept. 2024), p. 100572. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2024.100572. (Visited on 11/17/2025).

[38] Huy Hoang Nguyen et al. *Variational Autoencoder for Anomaly Detection: A Comparative Study*. https://arxiv.org/abs/2408.13 Aug. 2024. (Visited on 11/13/2025).

[39] Ethan Perez et al. *FiLM: Visual Reasoning with a General Conditioning Layer*. Dec. 2017. DOI: 10.48550/arXiv.1709.07871. arXiv: 1709.07871 [cs]. (Visited on 03/25/2026).

[40] Erwin Platen. "D1.2 Report on Anomaly Detection Feasibility and Methods Envisaged". In: (Oct. 2024), p. 82.

[41] Erwin Platen. "D2.3 Software Design Document". In: ().

[42] Nathalie Poulet. *Megha-Tropiques Technical Memorandum n° 3: Quality of geostationary satellite images – Megha-Tropiques*. Dec. 2017. (Visited on 03/26/2026).

[43] Soumi Ray et al. "Using Statistical Anomaly Detection Models to Find Clinical Decision Support Malfunctions". In: *Journal of the American Medical Informatics Association* 25.7 (July 2018), pp. 862–871. ISSN: 1527-974X. DOI: 10.1093/jamia/ocy041. (Visited on 05/01/2026).

[44] I.S. Reed and X. Yu. "Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.10 (Oct. 1990), pp. 1760–1770. ISSN: 0096-3518. DOI: 10.1109/29.60107. (Visited on 11/18/2025).

[45] Blaž Rolih et al. "Divide and Conquer: High-Resolution Industrial Anomaly Detection via Memory Efficient Tiled Ensemble". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2024, pp. 3866–3875. ISBN: 979-8-3503-6547-4. DOI: 10.1109/CVPRW63382.2024.00391. (Visited on 01/22/2026).

[46] Lukas Ruff et al. *A Unifying Review of Deep and Shallow Anomaly Detection*. https://arxiv.org/abs/2009.11732v3. Sept. 2020. DOI: 10.1109/JPROC.2021.3052449. (Visited on 06/05/2026).

[47] Mayu Sakurada and Takehisa Yairi. "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction". In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. MLSDA'14. New York, NY, USA: Association for Computing Machinery, Dec. 2014, pp. 4–11. ISBN: 978-1-4503-3159-3. DOI: 10.1145/2689746.2689747. (Visited on 11/17/2025).

[48] Thomas Schlegl et al. "F-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks". In: *Medical Image Analysis* 54 (May 2019), pp. 30–44. ISSN: 1361-8415. DOI: 10.1016/j.media.2019.01.010. (Visited on 11/17/2025).

[49] Thomas Schlegl et al. *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*. Mar. 2017. DOI: 10.48550/arXiv.1703.05921. arXiv: 1703.05921 [cs]. (Visited on 12/04/2025).

[50] Johannes Schmetz et al. "AN INTRODUCTION TO METEOSAT SECOND GENERATION (MSG)". In: *Bulletin of the American Meteorological Society* 83.7 (July 2002), pp. 977–992. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(2002)083<0977:AITMSG>2.3.CO;2. (Visited on 05/21/2026).

[51] Bernhard Schölkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: *Neural Computation* 13.7 (2001), pp. 1443–1471. ISSN: 0899-7667. DOI: 10.1162/089976601750264965.

[52] Jacob Senior. "D1.7 Feasibility Report of the Machine Learning Techniques in GIAD". In: (June 2022), p. 29.

[53] Xinyu Sheng, Shande Tuo, and Lu Wang. "Surface Anomaly Detection and Localization with Diffusion-based Reconstruction". In: *2024 International Joint Conference on Neural Networks (IJCNN)*. June 2024, pp. 1–8. DOI: 10.1109/IJCNN60899.2024.10651406. (Visited on 05/01/2026).

[54] Yongju Son et al. "LSTM–GAN Based Cloud Movement Prediction in Satellite Images for PV Forecast". In: *Journal of Ambient Intelligence and Humanized Computing* 14.9 (Sept. 2023), pp. 12373–12386. ISSN: 1868-5145. DOI: 10.1007/s12652-022-04333-7. (Visited on 05/19/2026).

[55] *Spectral Band - an Overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/engineering/spectral-band. (Visited on 05/21/2026).

[56] E. G. Stassinopoulos et al. "Radiation-Induced Anomalies in Satellites". In: *Journal of Spacecraft and Rockets* 33.6 (Nov. 1996), pp. 877–882. ISSN: 0022-4650. DOI: 10.2514/3.26854. (Visited on 05/13/2026).

[57] *The Basics of Spectral Bands - NIR, SWIR, and RGB*. https://swiftgeospatial.solutions/2025/03/12/the-basics-of-spectral-bands-nir-swir-and-rgb/. (Visited on 05/21/2026).

[58] Mayur V. Tiwari et al. “The Analysis of Satellite Images for Anomaly Identification in Patterns”. In: *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*. Feb. 2025, pp. 159–163. DOI: 10.1109/ICCCIT62592.2025.10928158. (Visited on 11/13/2025).

[59] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. “A Survey of Transfer Learning”. In: *Journal of Big Data* 3.1 (May 2016), p. 9. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6. (Visited on 11/18/2025).

[60] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. *Towards High-Resolution Industrial Image Anomaly Detection*. Aug. 2025. DOI: 10.48550/arXiv.2508.12931. arXiv: 2508.12931 [cs]. (Visited on 01/22/2026).

[61] Chong Zhou and Randy C. Paffenroth. “Anomaly Detection with Robust Deep Autoencoders”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 665–674. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098052. (Visited on 11/17/2025).

[62] Guangxuan Zhu et al. “A Novel LSTM-GAN Algorithm for Time Series Anomaly Detection”. In: *2019 Prognostics and System Health Management Conference (PHM-Qingdao)*. Oct. 2019, pp. 1–6. DOI: 10.1109/PHM-Qingdao46334.2019.8942842. (Visited on 05/19/2026).

[63] Fuzhen Zhuang et al. “A Comprehensive Survey on Transfer Learning”. In: *Proceedings of the IEEE* 109.1 (Jan. 2021), pp. 43–76. ISSN: 1558-2256. DOI: 10.1109/JPROC.2020.3004555. (Visited on 11/18/2025).

A

Data Preparation Pipeline

This appendix describes the full pipeline used to convert raw geostationary satellite data into the tiled dataset used for training and evaluation. The pipeline consists of three stages: raw data ingestion and resizing, tiling, and test set construction.

A.1. Raw Data Ingestion and Resizing

Raw satellite images are read from disk via the GIAD plugin system, which handles the instrument-specific file formats for each satellite, using the dataloaders provided by the satellite operators. Three satellites are included, with the channels listed in Table A.1.

Satellite	**Channels**
MET09 (MSG)	VIS0060, VIS0080, IR0160, IR0390, IR0870, IR0970, IR1080, IR1200, IR1340, WV0620, WV0730
MTSAT1	VIS0072, IR0375, WV0675, IR1080, IR1200
GOES11	VIS0065, IR0390, WV0670, IR1070, IR1190

Table A.1: Satellites and spectral channels included in the dataset.

Each image is loaded as a float32 array and normalised to $[0, 1]$ using the global min and max of that image. The images are then resized to 1024×1024 pixels using Lanczos resampling, which is better at preserving fine details than simpler methods like bilinear interpolation. For this dataset, only the full disk images are used, so rapid-scan images, only scanning parts of the earth are discarded. The loaded MSG images contain backscan data, which is cropped out to retain only the earth disk. The normalised image is scaled back to the uint16 range $[0, 65535]$ and saved as a 16-bit PNG with the output path `{satellite}/{channel}/{filename}.png`. GOES11 images are excluded from the baseline training set and reserved as the transfer learning target domain.

A.2. Tiling

The resized 1024×1024 images are split into 128×128 pixel tiles. Each image is first normalised to 8-bit by dividing by the full-image maximum (65535 for 16-bit images, 255 for 8-bit images) and scaling to $[0, 255]$. Using the full-image maximum rather than a per-tile maximum ensures that all tiles from the same image share the same intensity scale.

Tiles are extracted with a configurable stride; the default configuration uses non-overlapping tiles (stride = tile size). Tiles whose pixel values are entirely zero are discarded, as these correspond to regions outside the earth disk where no sensor data is present. The remaining tiles are saved to `{satellite}/{channel}/train/` with filenames encoding the row and column position (`{image}_tile_r{row}_c{col}.png`), which are used by the metadata encoder to derive the tile's spatial position within the image.

A.3. Train and Validation Split

The training tiles are split into a training set (90%) and a validation set (10%) by random sampling with a fixed seed (42), ensuring reproducibility. The dataloader applies a standard normalisation transform to map pixel values from [0, 255] to [−1, 1]:

$$x_{\text{norm}} = \frac{x/255 - 0.5}{0.5} \tag{A.1}$$

All-zero tiles are filtered out at load time by the collation function.

A.4. Test Set Construction

For each satellite channel, 100 full-resolution images are randomly sampled to form the test set. These are saved separately from the training tiles to prevent data leakage. The test set is divided into normal and anomalous subsets.

Normal images. The sampled images are normalised to 8-bit and saved to `test/normal/` without modification.

Anomalous images. A second sample of 100 images is drawn and augmented with synthetic anomalies. Six anomaly types are used, each applied to an equal share of the images:

Lines A block of 16–32 consecutive horizontal lines is zeroed out at a random vertical position.

Patches One to three rectangular or elliptical black patches (50–150 px) are placed at random positions.

Noise Two to five regions (50–150 px) are filled with uniform random pixel values.

Border black patches Small black patches (25–68 px) are placed at tile-boundary coordinates, to test the model's ability to handle anomalies that are on the edge of tiles.

Celestial body A single bright elliptical patch is placed at a random location, simulating a star or moon in the field of view.

Half-swap The left and right, or the bottom and top halves of the image are swapped, simulating a severe misalignment anomaly.

Each anomalous image is saved to `test/anomalous/{type}/` alongside a binary mask in `test/anomalous/masks/{ty` where mask pixels are 255 at anomaly locations and 0 elsewhere. These masks are used for pixel-level localisation evaluation.

Fine-grain anomalies. A second set of anomalous images is created using the same anomaly types but with significantly smaller anomaly sizes (roughly 10× smaller in linear extent). These are saved to `test/fine_grain_anomalies` and used to evaluate sensitivity to subtle anomalies.

B

GIAD Rule-Based Detection Algorithms

This appendix describes the rule-based anomaly detectors in the production GIAD pipeline, providing context for the comparison with the learned detector in Section 6.7.

B.1. Overview

An overview of the GIAD rule-based detection pipeline is shown in Figure B.1.

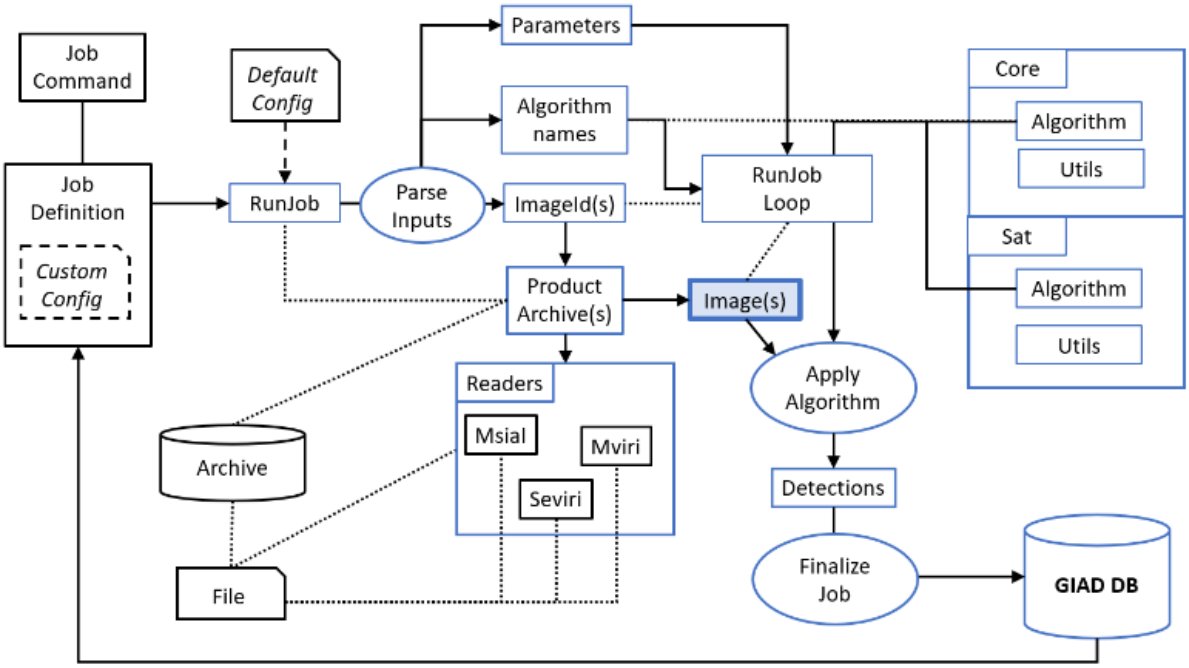


Figure B.1: Schematic of the GIAD rule-based detection pipeline.

Satellite images are first ingested by instrument-specific readers, which decode the raw data into a common image format. The GIAD plugin system then identifies the satellite and channel being processed and applies the appropriate set of detectors. The outputs of all active detectors are aggregated into a final anomaly mask for the image.

B.2. Detector Descriptions

The GIAD pipeline contains a number of detectors, each targeting a specific anomaly type such as line artefacts, hot pixels, or spatial misalignment. The pipeline is actively developed, so not all detectors are described

here in full; instead, the common design principles are outlined.

All detectors follow a shared design: each is implemented as a standalone function with explicit arguments and documented default values, applies anomaly-type-specific detection logic, and returns a standardised output tuple. This ensures that detectors are independently testable and consistent in the information they report to the pipeline.

For a more indepth description of the exact algorithms and the detection pipeline, please refer to the GIAD documentation Hogervorst and Senior-Williams [18], Platen [41], and Hogervorst [17]

C

Extended Visualisation Results

This appendix provides additional reconstruction and heatmap examples for the baseline, overlap, and metadata experiments, supplementing the selected figures in Chapter 5. Each image shows the original input, tile-wise reconstruction, pixel-error map, and combined anomaly heatmap. The filename encodes the maximum per-pixel anomaly score (`px_max`) and the classification threshold (`thr`) used for that experiment.

C.1. Baseline Results

C.1.1. True Positives

Anomalous images correctly classified as anomalous (px_max > threshold = 0.09).

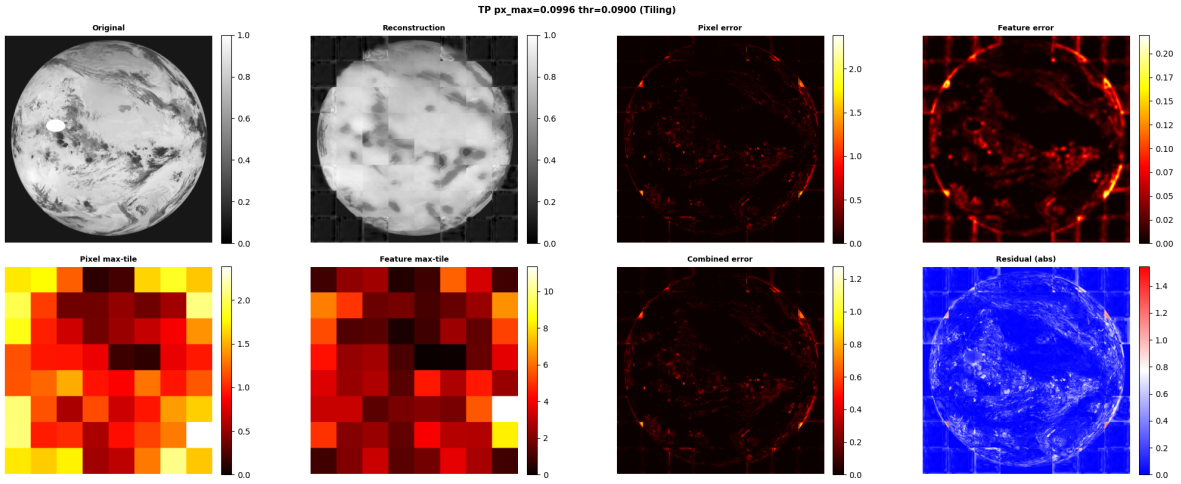


Figure C.1: True positive — px_max = 0.0996, threshold = 0.09 (detection 008).

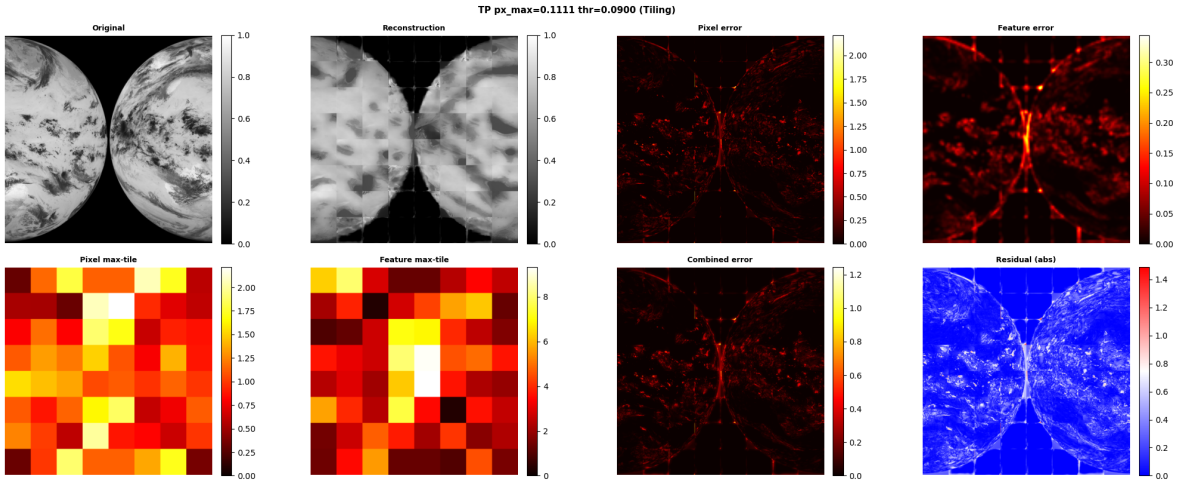


Figure C.2: True positive — px_max = 0.1111, threshold = 0.09 (detection 020).

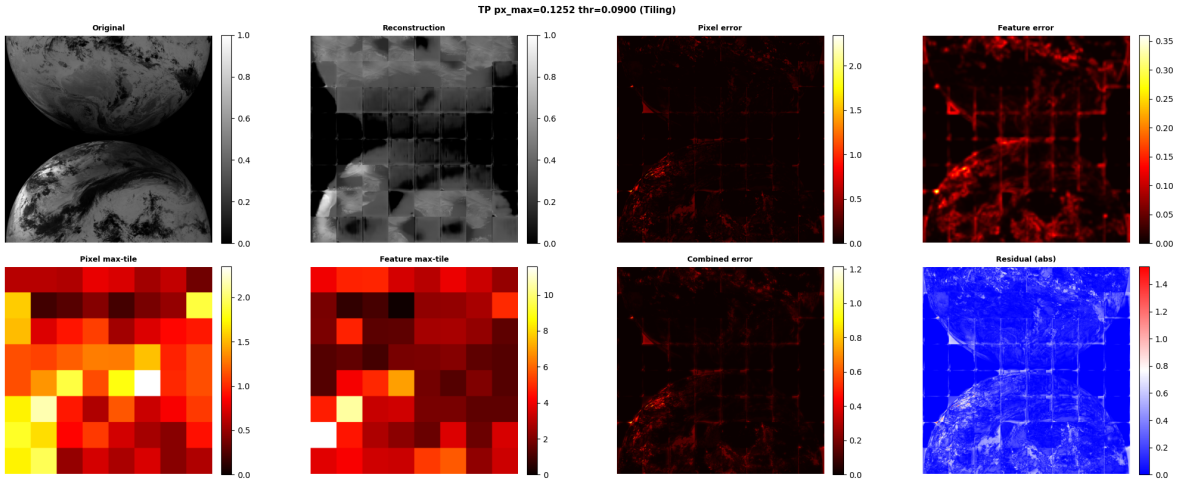


Figure C.3: True positive — px_max = 0.1252, threshold = 0.09 (detection 017).

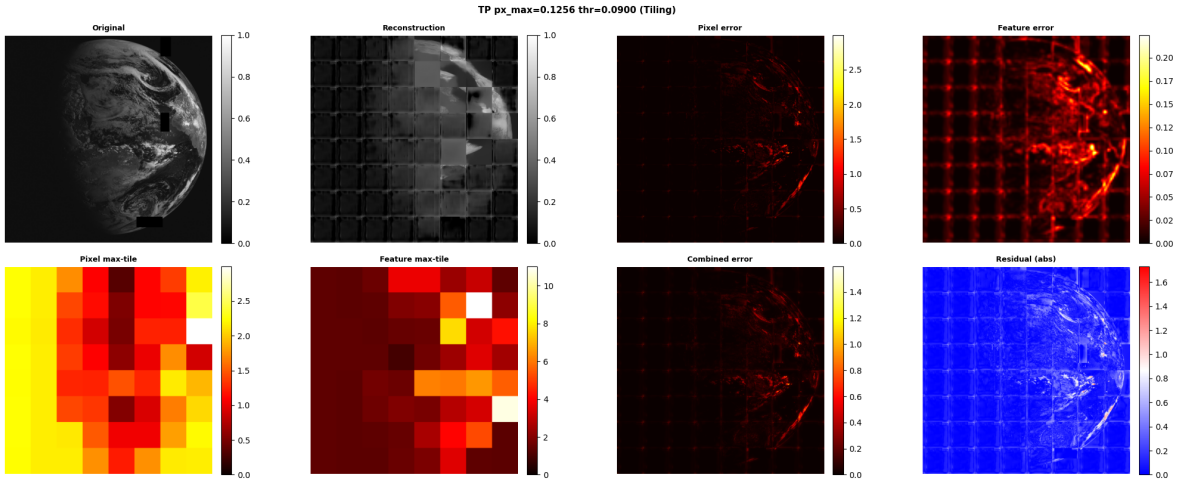


Figure C.4: True positive — px_max = 0.1256, threshold = 0.09 (detection 012).

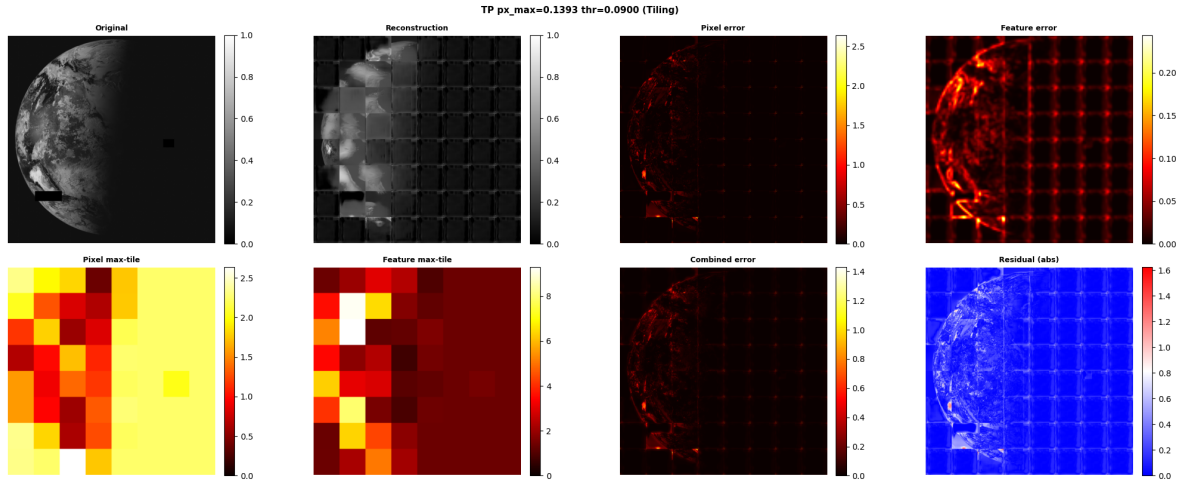


Figure C.5: True positive — px_max = 0.1393, threshold = 0.09 (detection 000).

C.1.2. True Negatives

Normal images correctly classified as normal (px_max ≤ threshold = 0.09).

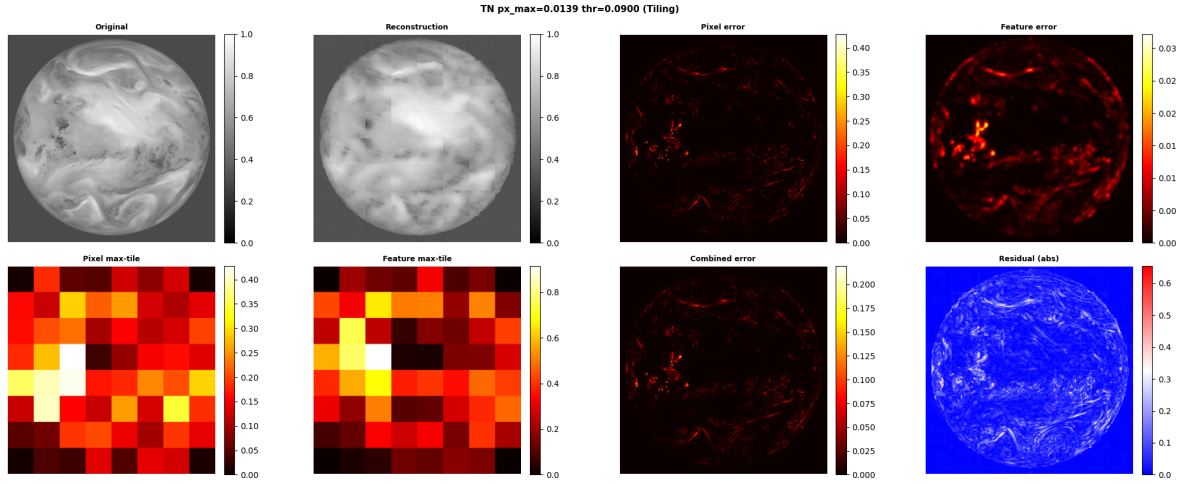


Figure C.6: True negative — px_max = 0.0139, threshold = 0.09 (detection 014).

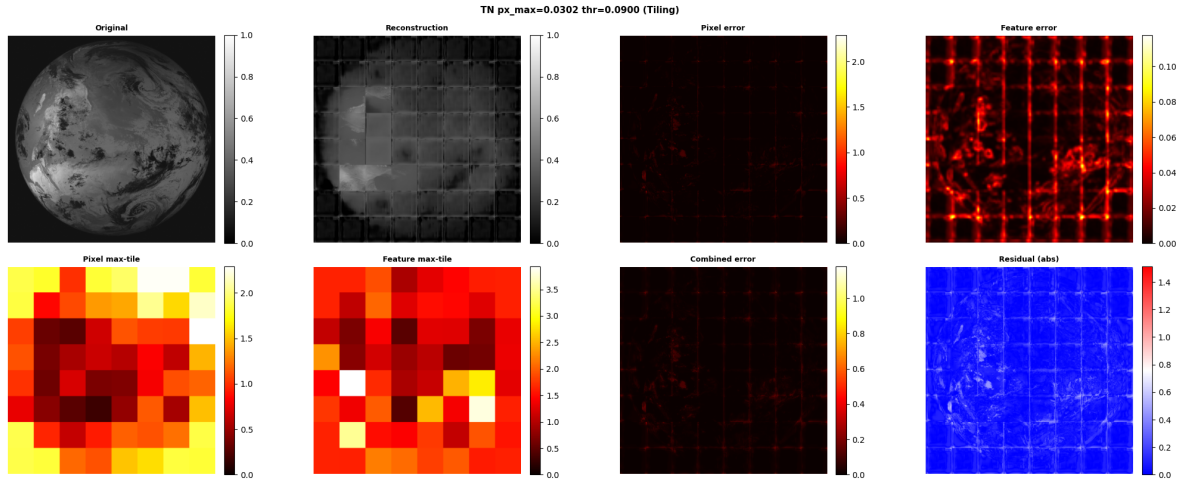


Figure C.7: True negative — px_max = 0.0302, threshold = 0.09 (detection 003).

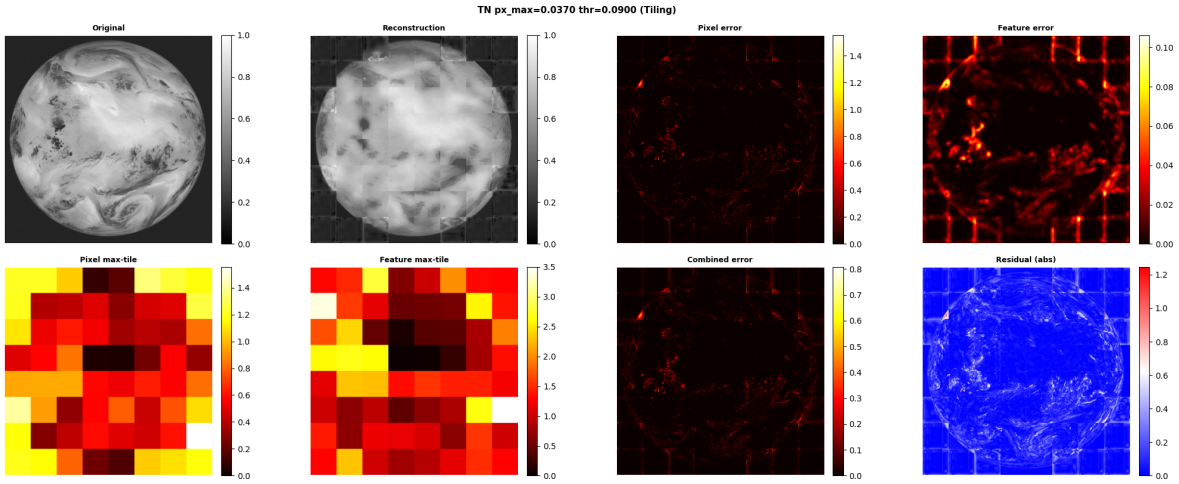


Figure C.8: True negative — px_max = 0.0370, threshold = 0.09 (detection 016).

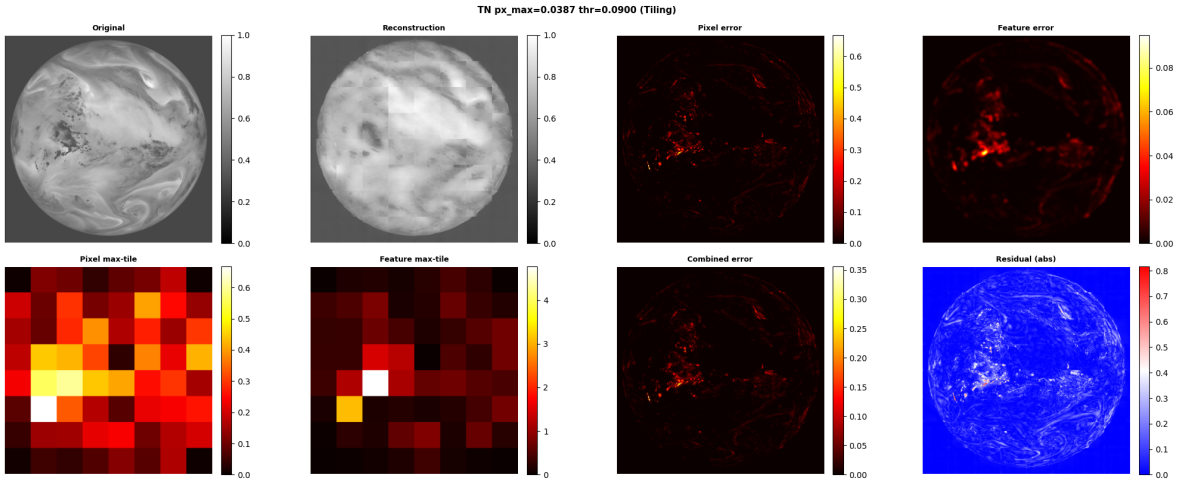


Figure C.9: True negative — px_max = 0.0387, threshold = 0.09 (detection 015).

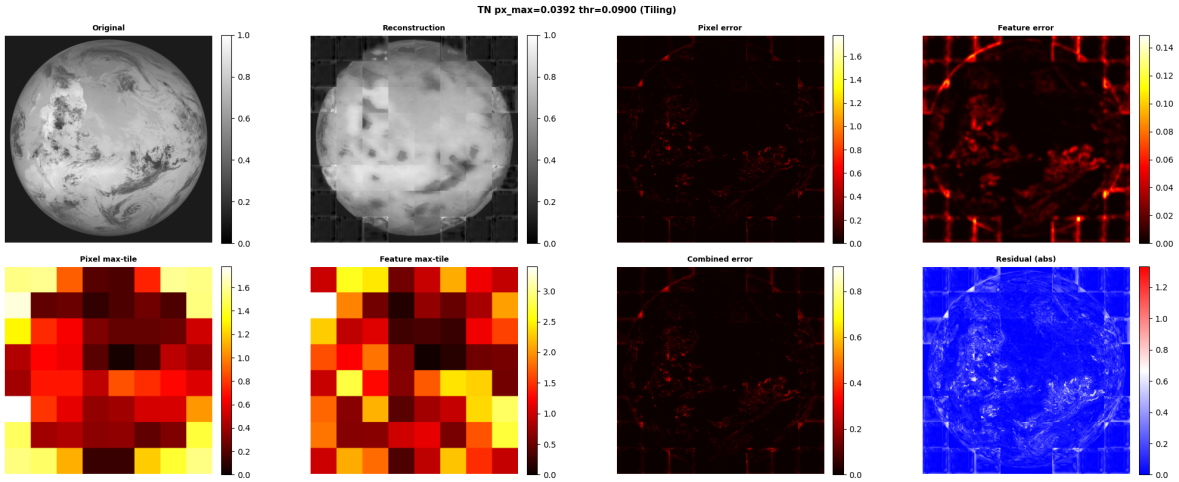


Figure C.10: True negative — px_max = 0.0392, threshold = 0.09 (detection 006).

C.1.3. False Positives

Normal images incorrectly classified as anomalous (px_max > threshold = 0.09).

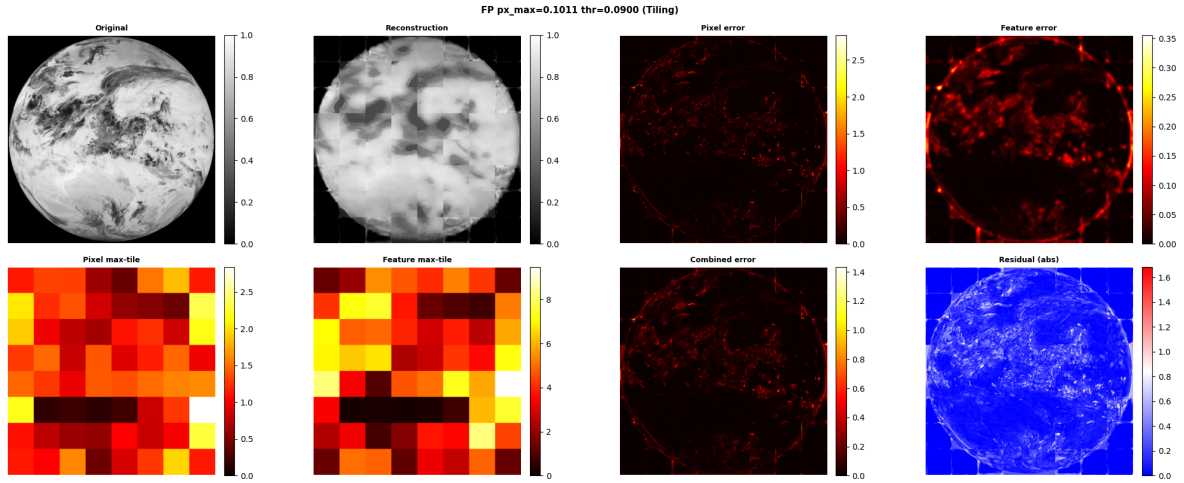


Figure C.11: False positive — px_max = 0.1011, threshold = 0.09 (detection 020).

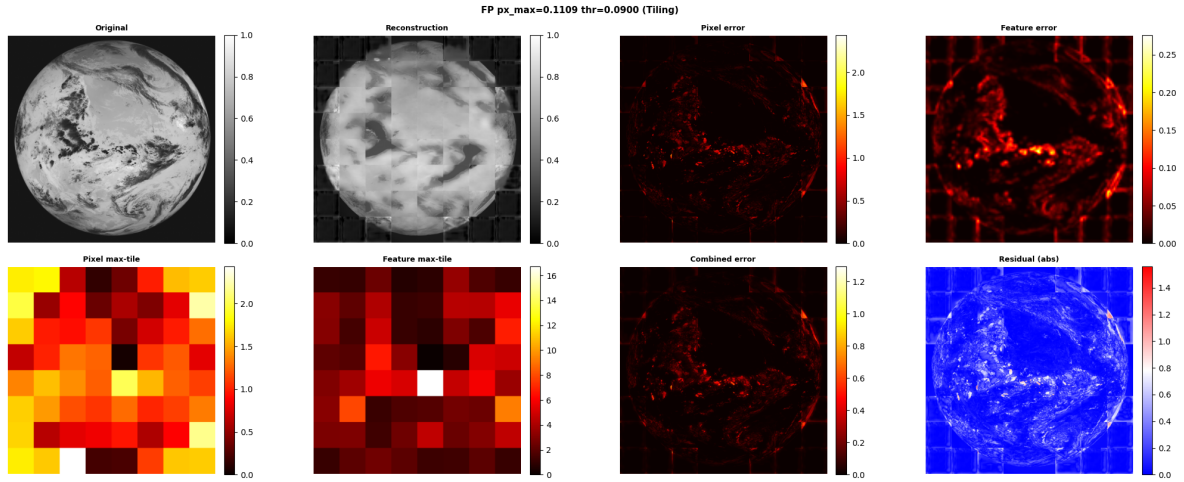


Figure C.12: False positive — px_max = 0.1109, threshold = 0.09 (detection 004).

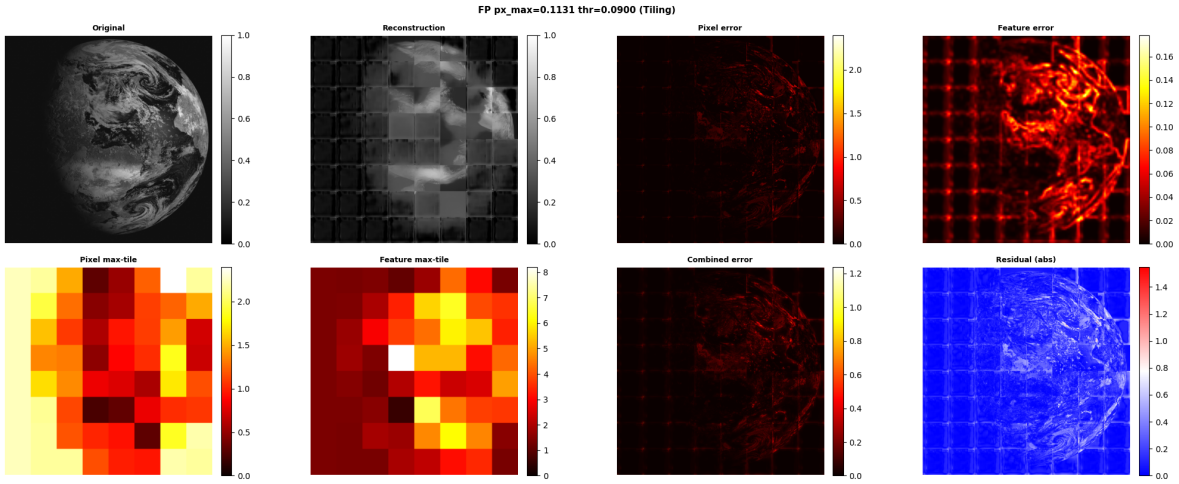


Figure C.13: False positive — px_max = 0.1131, threshold = 0.09 (detection 000).

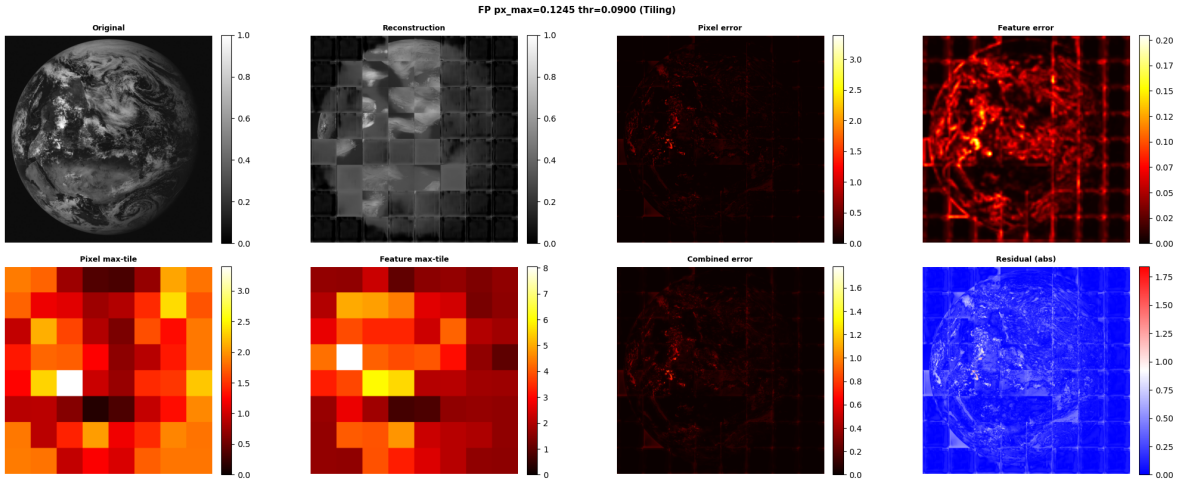


Figure C.14: False positive — px_max = 0.1245, threshold = 0.09 (detection 011).

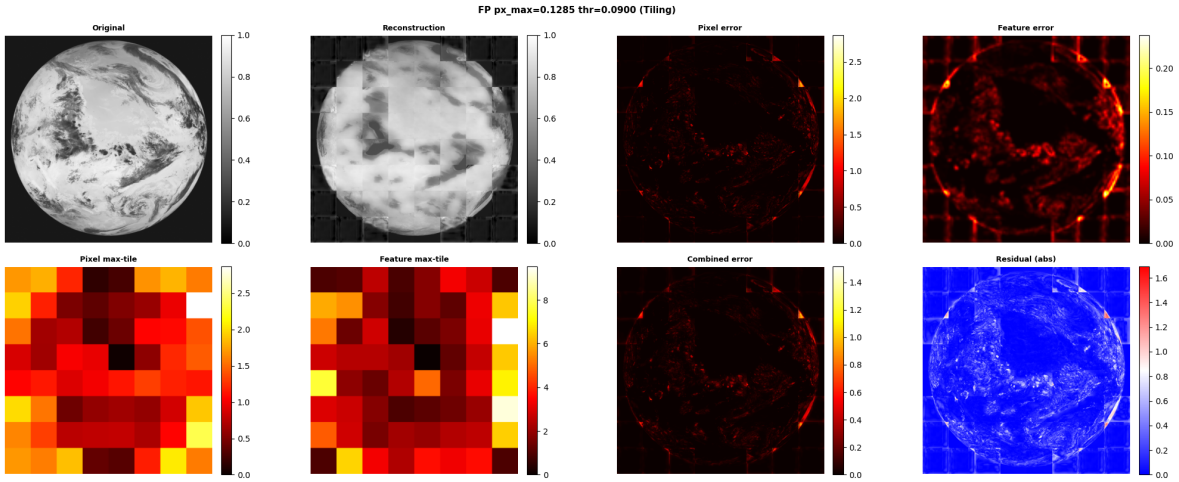


Figure C.15: False positive — px_max = 0.1285, threshold = 0.09 (detection 009).

C.1.4. False Negatives

Anomalous images incorrectly classified as normal (px_max ≤ threshold = 0.09). All four baseline false negatives are shown.

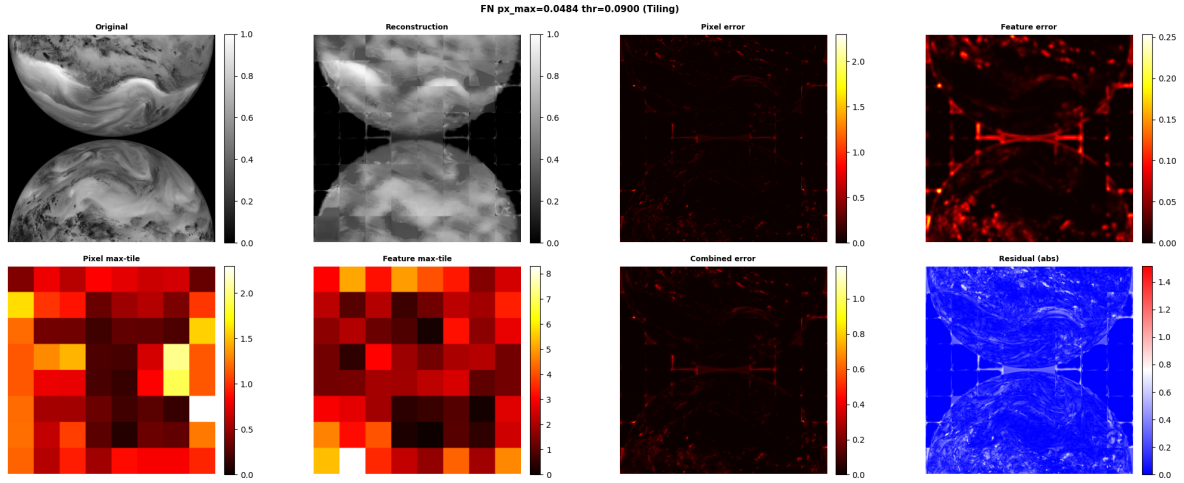


Figure C.16: False negative — px_max = 0.0484, threshold = 0.09 (detection 023).

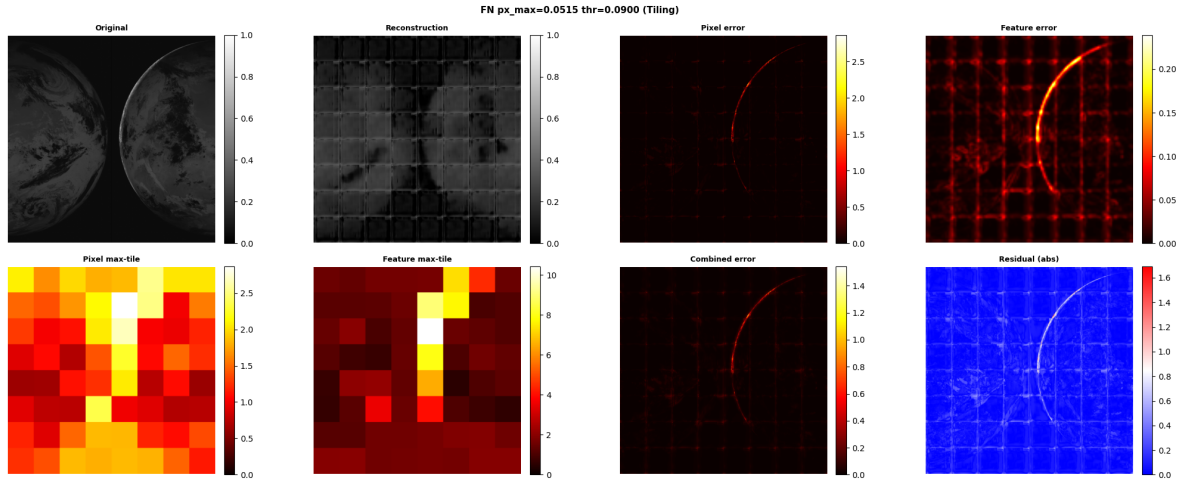


Figure C.17: False negative — px_max = 0.0515, threshold = 0.09 (detection 002).

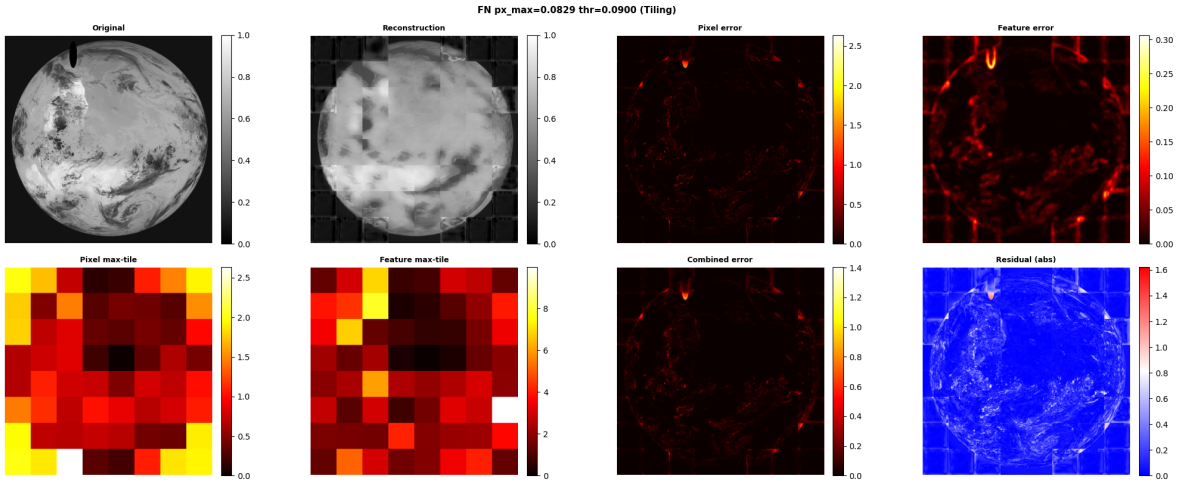


Figure C.18: False negative — px_max = 0.0829, threshold = 0.09 (detection 009).

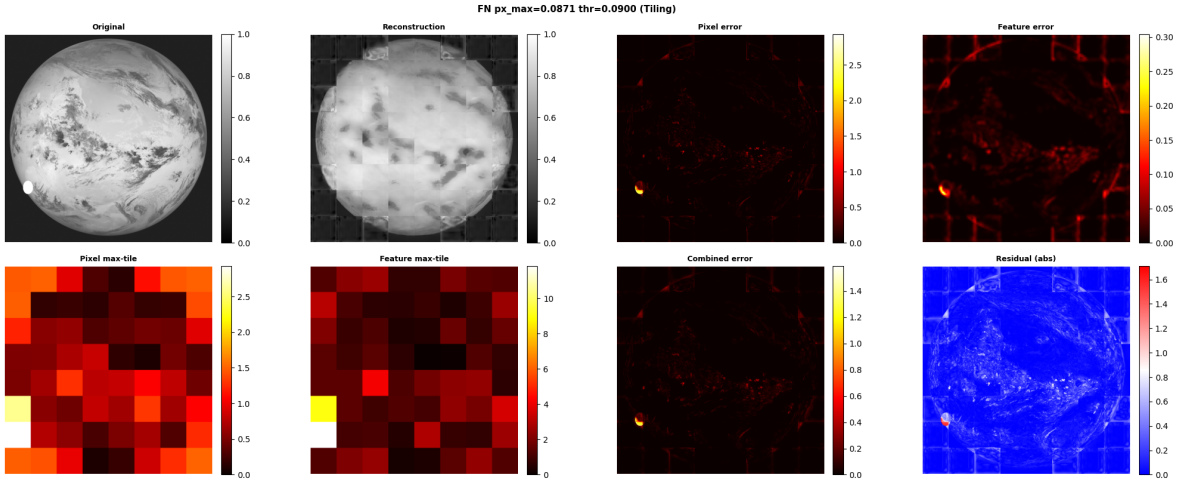


Figure C.19: False negative — px_max = 0.0871, threshold = 0.09 (detection 005).

C.2. Overlap Results

C.2.1. True Positives

Anomalous images correctly classified as anomalous (px_max > threshold = 0.0547).

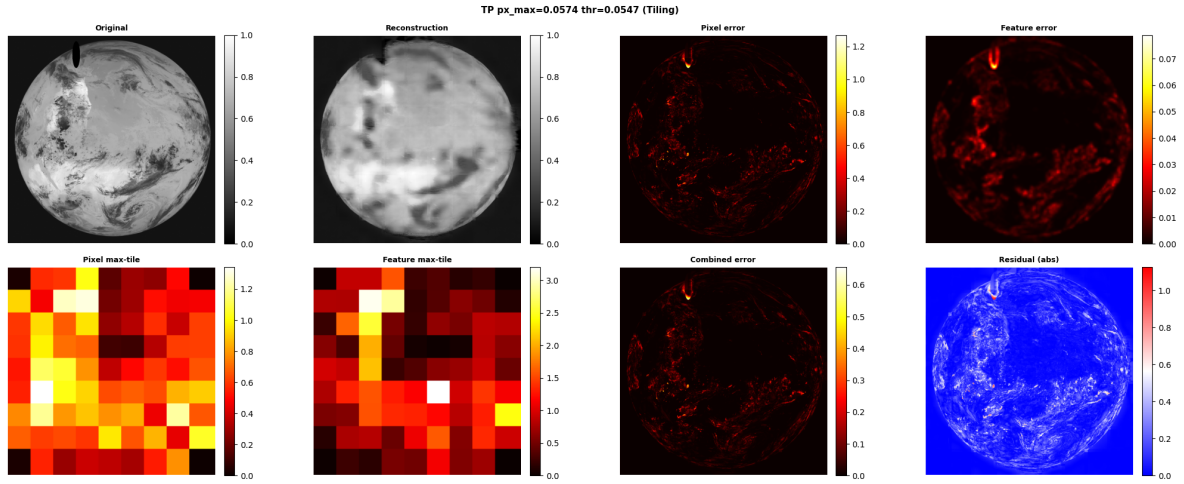


Figure C.20: True positive — px_max = 0.0574, threshold = 0.0547 (detection 009).

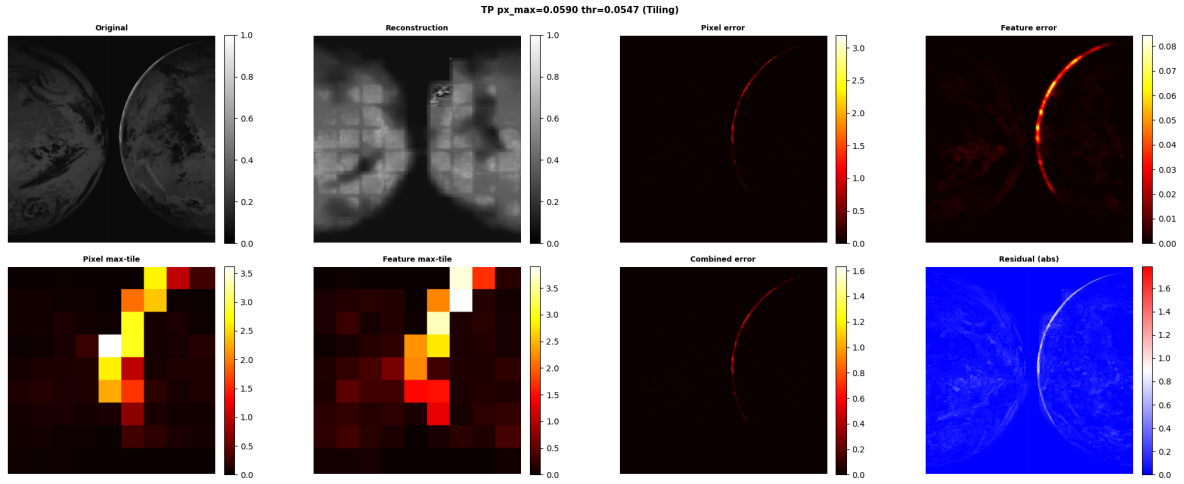


Figure C.21: True positive — px_max = 0.0590, threshold = 0.0547 (detection 002).

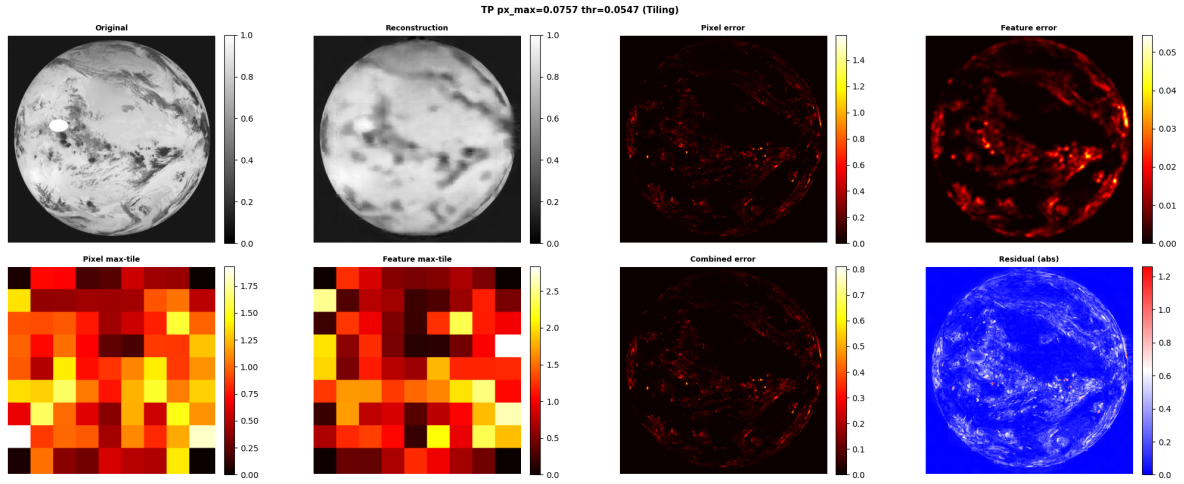


Figure C.22: True positive — px_max = 0.0757, threshold = 0.0547 (detection 008).

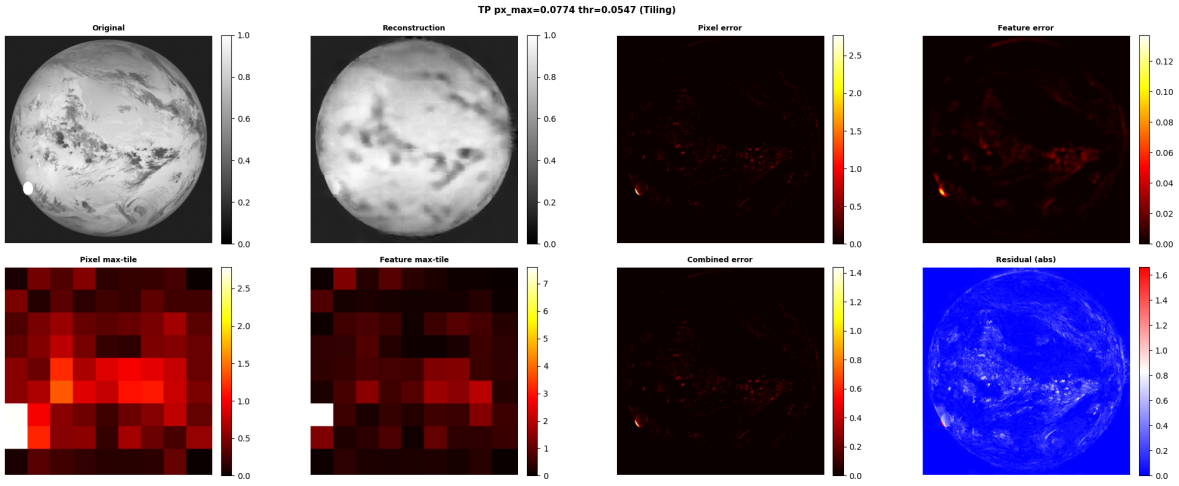


Figure C.23: True positive — px_max = 0.0774, threshold = 0.0547 (detection 005).

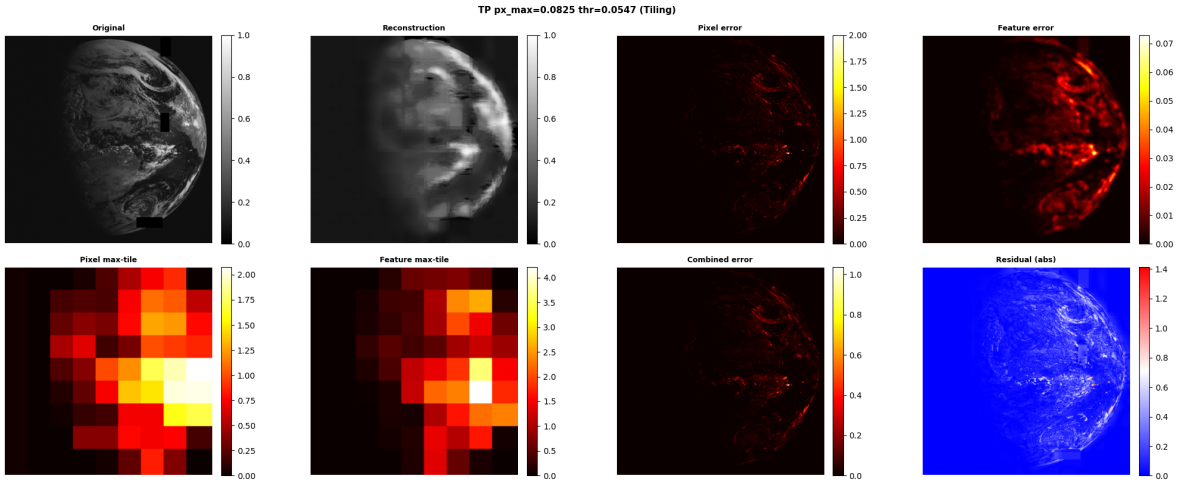


Figure C.24: True positive — px_max = 0.0825, threshold = 0.0547 (detection 012).

C.2.2. True Negatives

Normal images correctly classified as normal (px_max ≤ threshold = 0.0547).

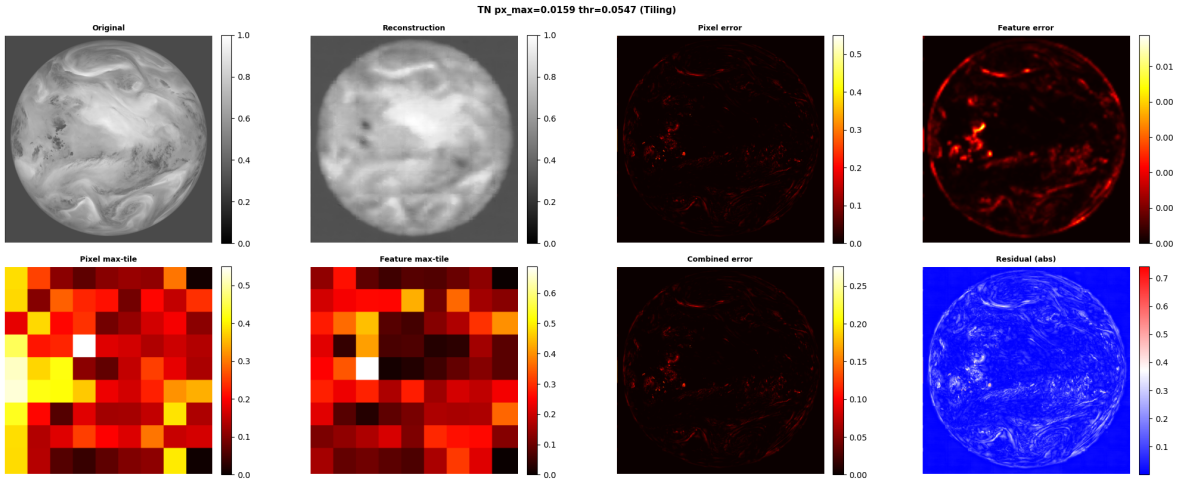


Figure C.25: True negative — px_max = 0.0159, threshold = 0.0547 (detection 014).

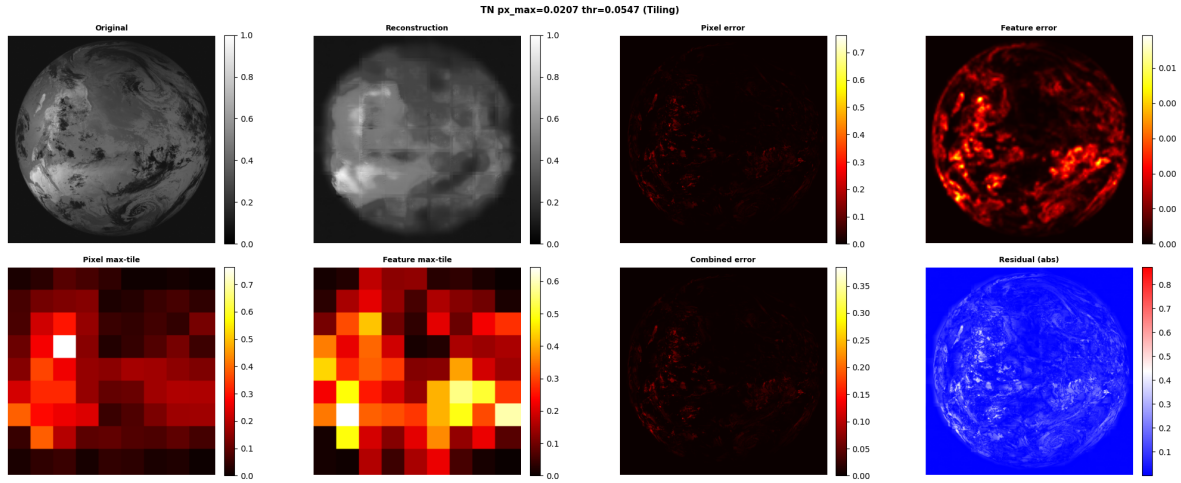


Figure C.26: True negative — px_max = 0.0207, threshold = 0.0547 (detection 003).

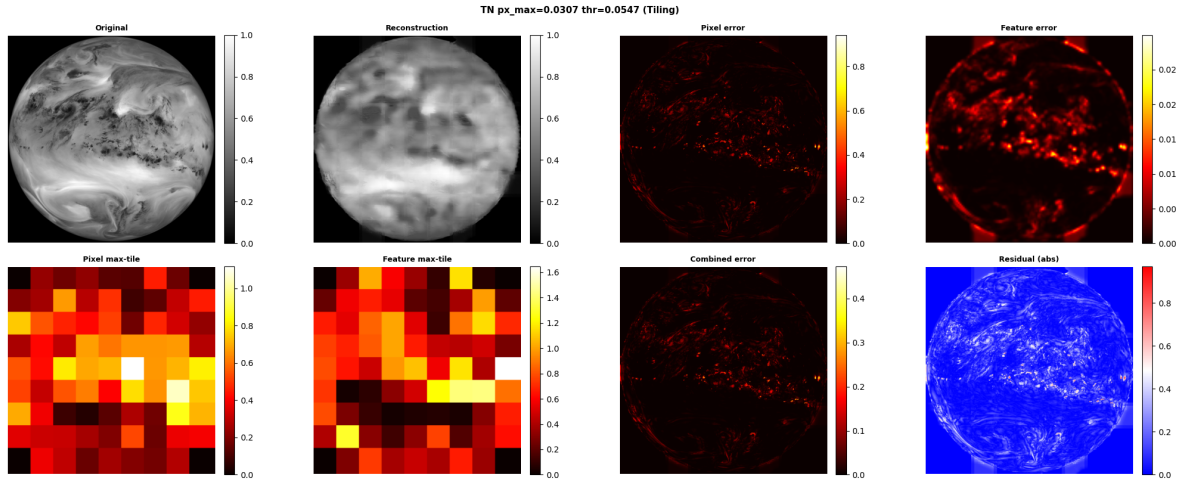


Figure C.27: True negative — px_max = 0.0307, threshold = 0.0547 (detection 023).

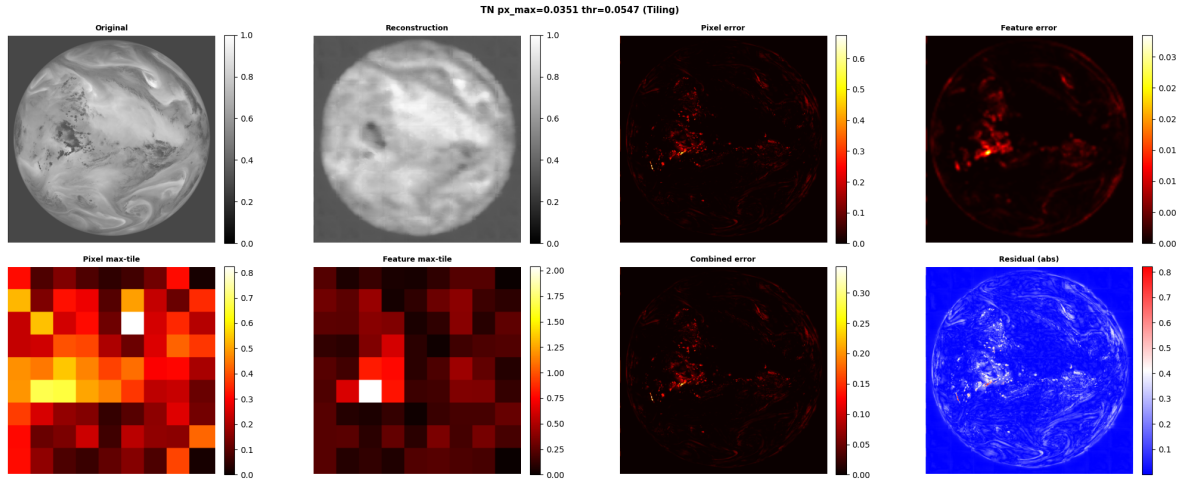


Figure C.28: True negative — px_max = 0.0351, threshold = 0.0547 (detection 015).

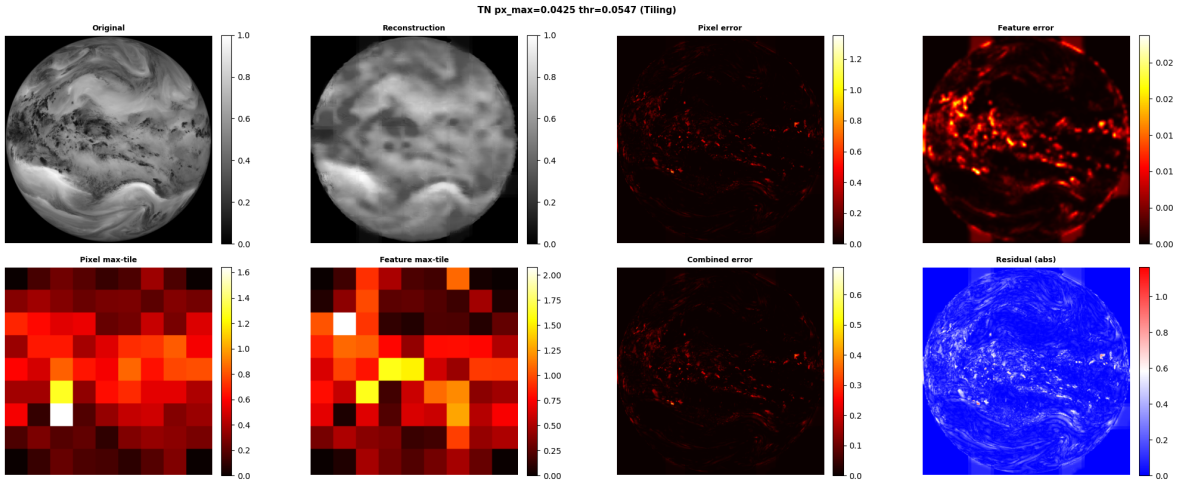


Figure C.29: True negative — px_max = 0.0425, threshold = 0.0547 (detection 024).

C.2.3. False Positives

Normal images incorrectly classified as anomalous (px_max > threshold = 0.0547).

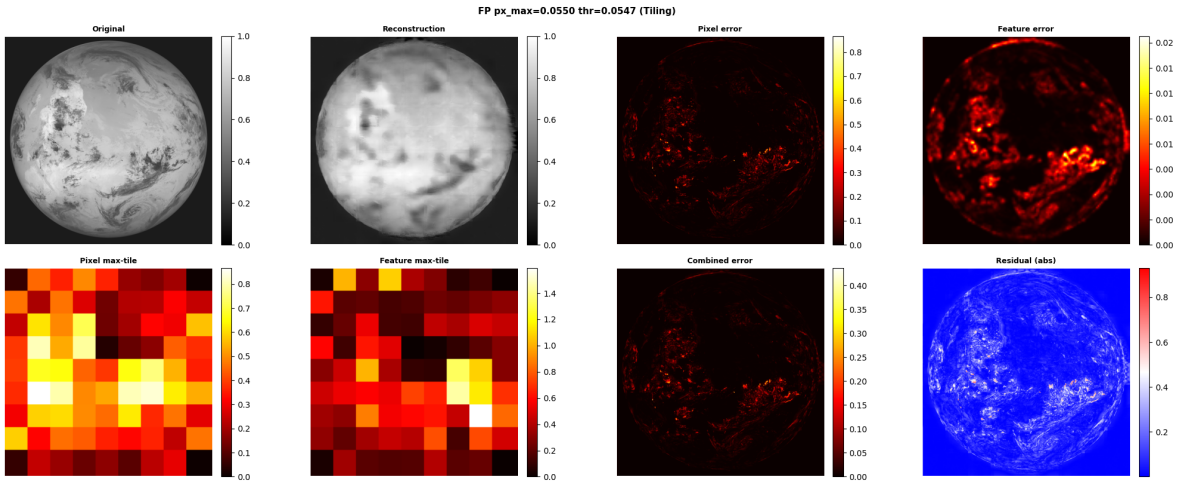


Figure C.30: False positive — px_max = 0.0550, threshold = 0.0547 (detection 006).

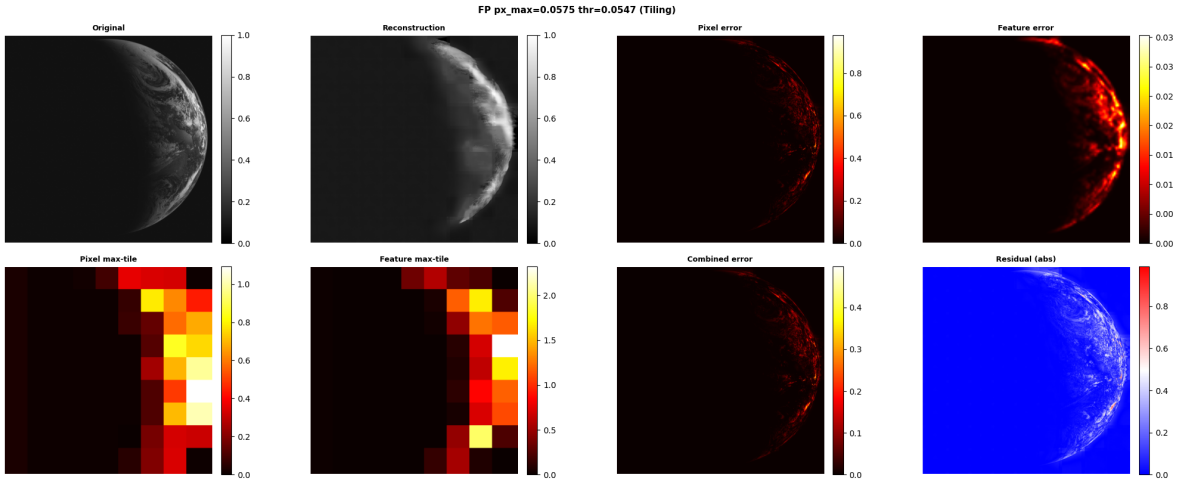


Figure C.31: False positive — px_max = 0.0575, threshold = 0.0547 (detection 012).

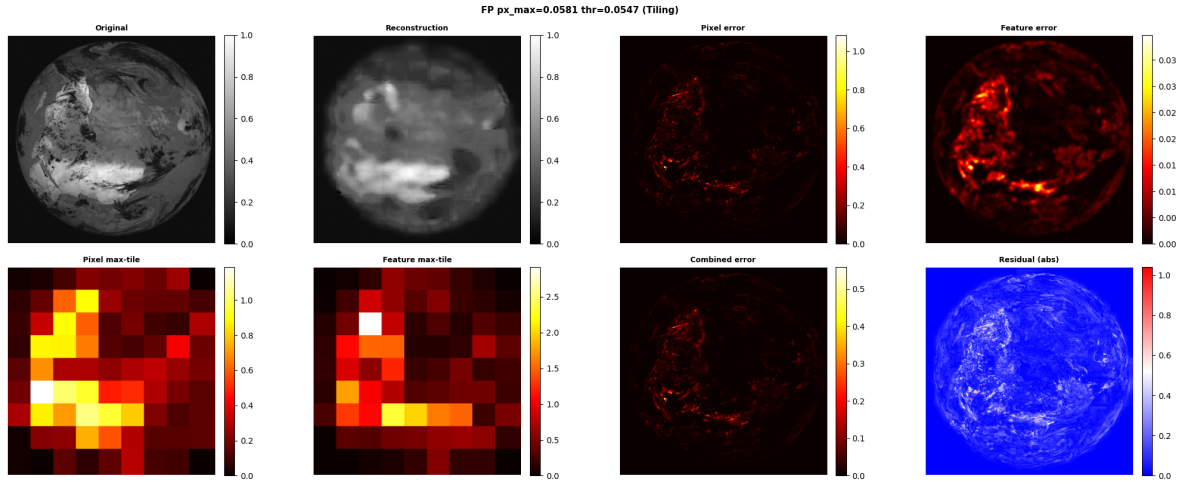


Figure C.32: False positive — px_max = 0.0581, threshold = 0.0547 (detection 002).

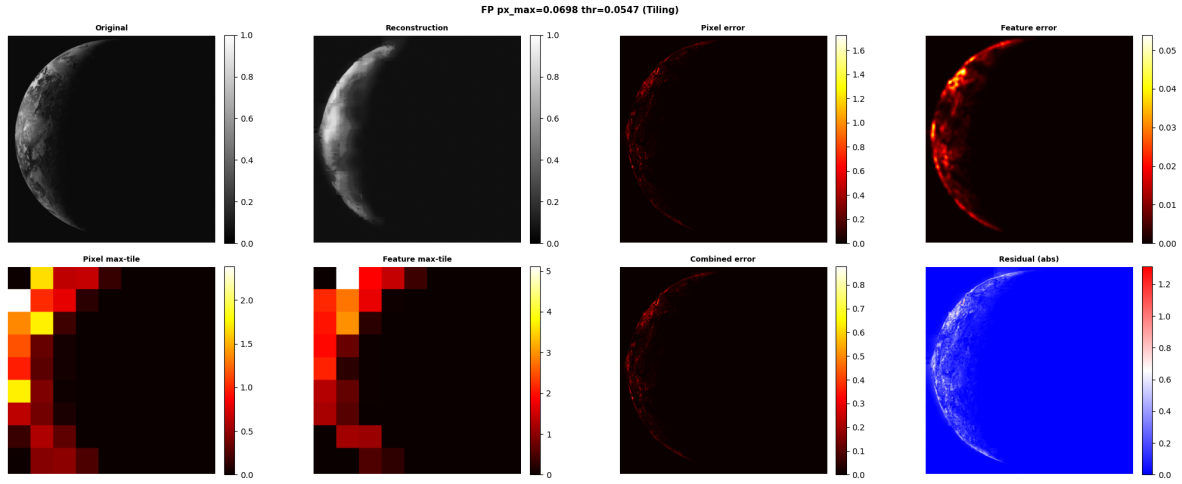


Figure C.33: False positive — px_max = 0.0698, threshold = 0.0547 (detection 001).

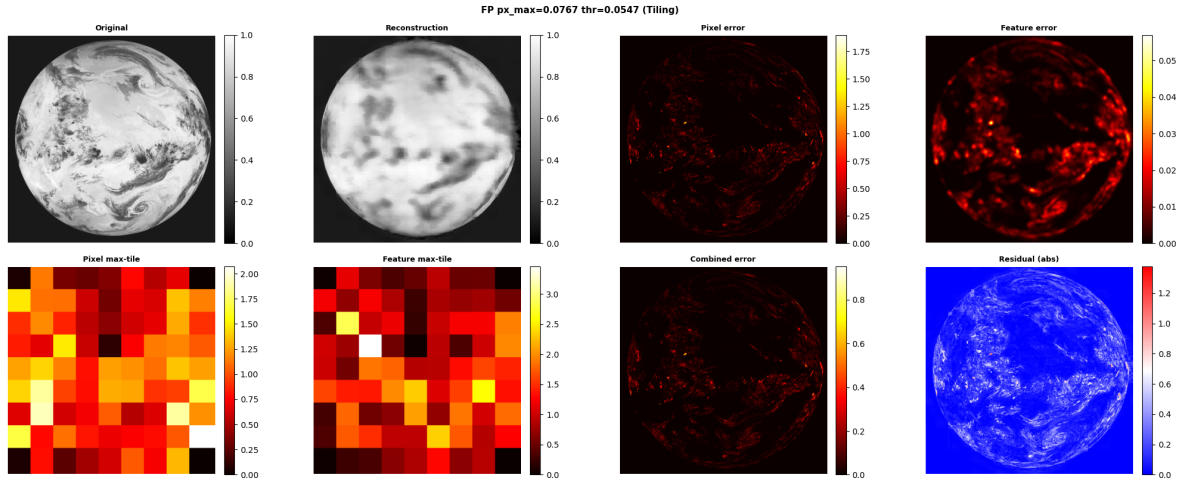


Figure C.34: False positive — px_max = 0.0767, threshold = 0.0547 (detection 008).

C.3. Metadata Conditioning Results

C.3.1. True Positives

Anomalous images correctly classified as anomalous (px_max > threshold = 0.0705).

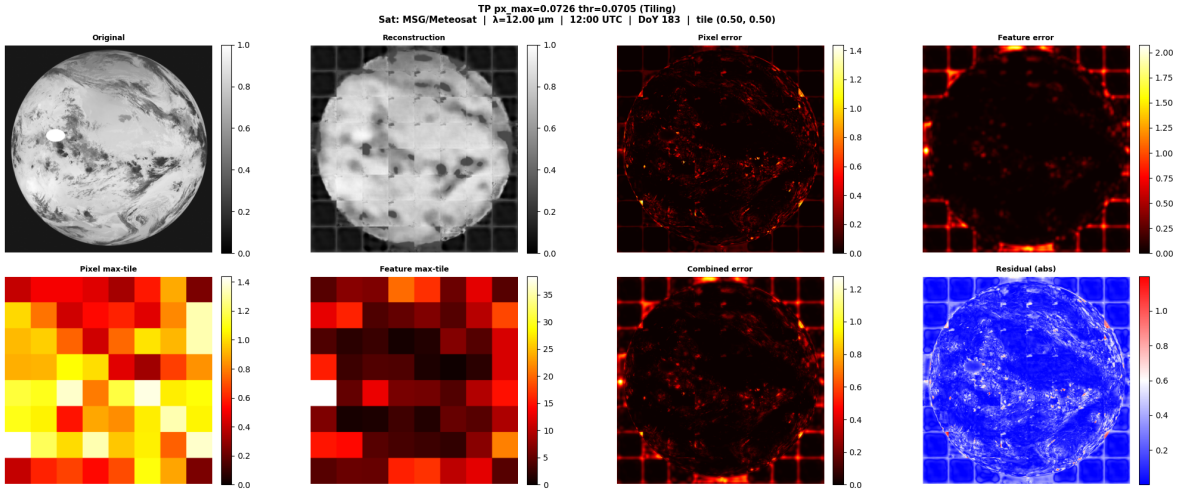


Figure C.35: True positive — px_max = 0.0726, threshold = 0.0705 (detection 008).

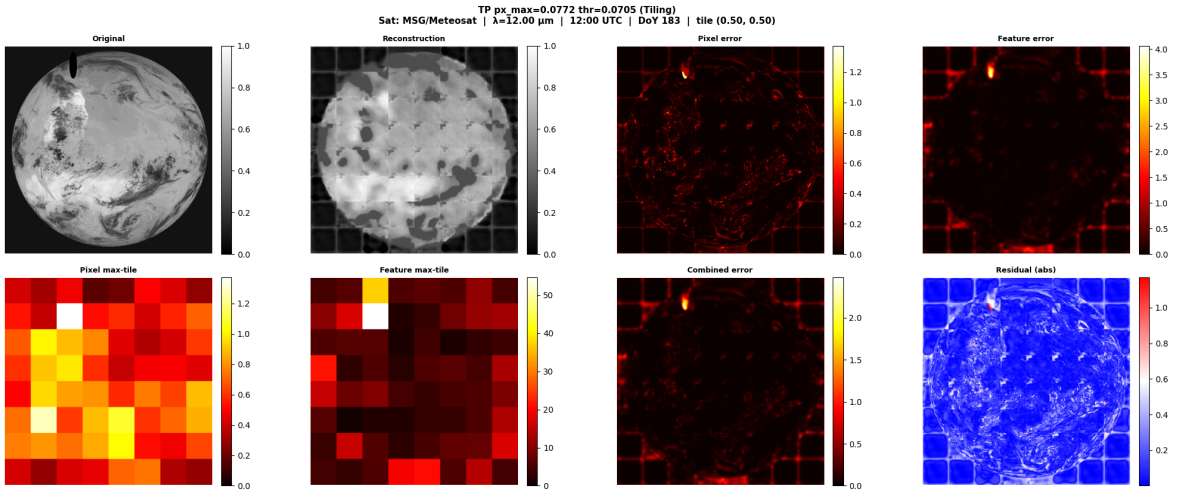


Figure C.36: True positive — px_max = 0.0772, threshold = 0.0705 (detection 009).

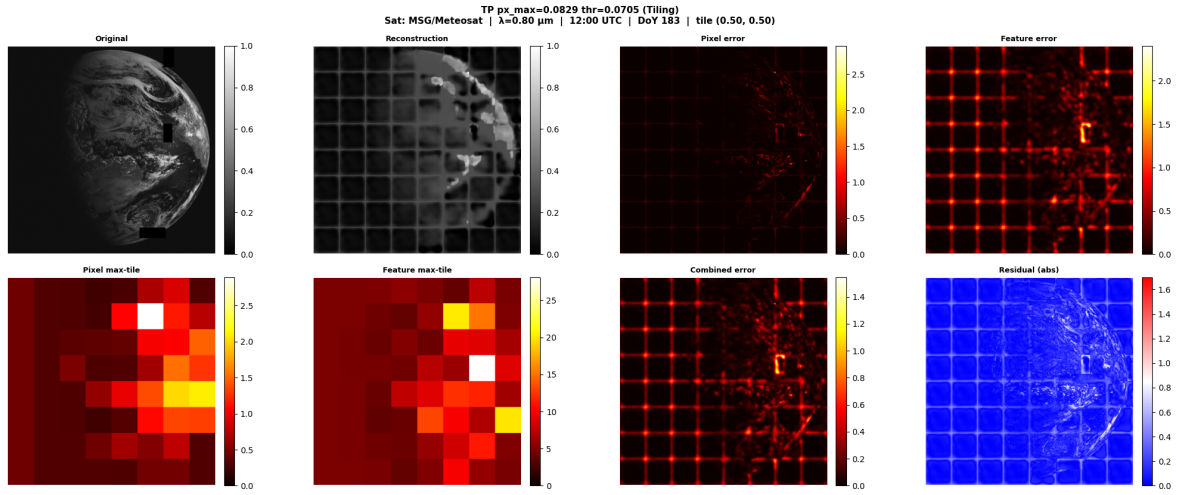


Figure C.37: True positive — px_max = 0.0829, threshold = 0.0705 (detection 012).

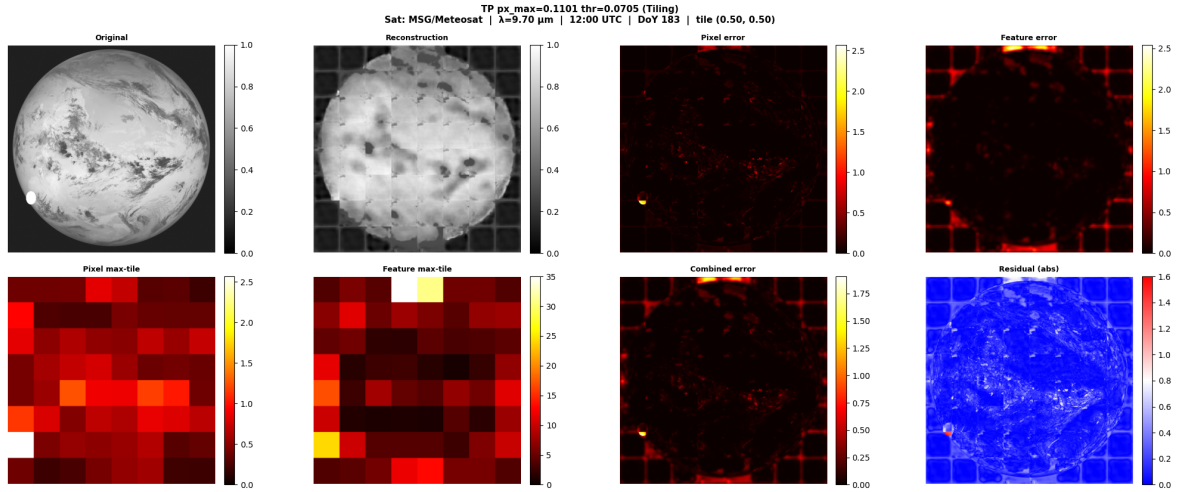


Figure C.38: True positive — px_max = 0.1101, threshold = 0.0705 (detection 005).

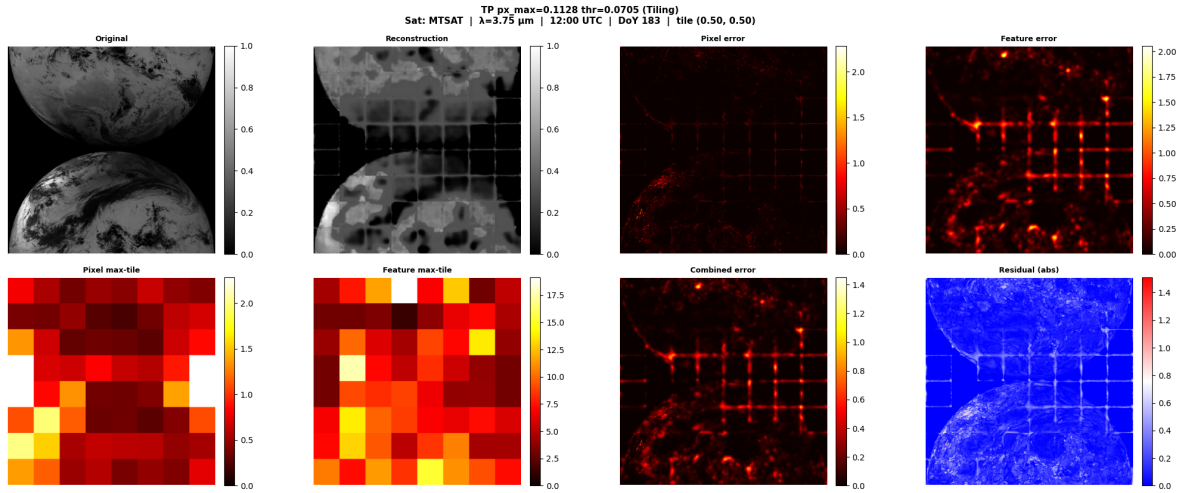


Figure C.39: True positive — px_max = 0.1128, threshold = 0.0705 (detection 017).

C.3.2. True Negatives

Normal images correctly classified as normal (px_max ≤ threshold = 0.0705).

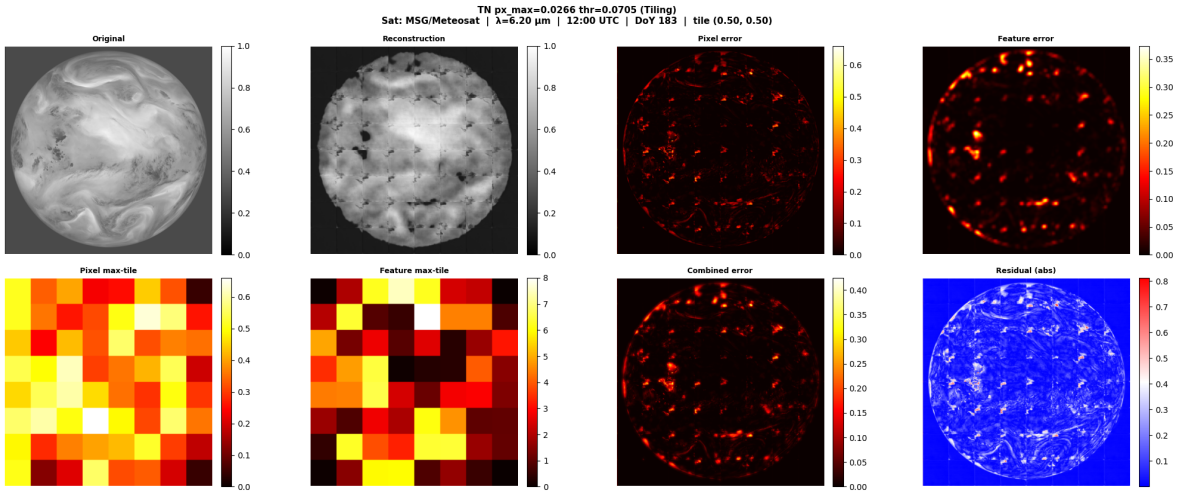


Figure C.40: True negative — px_max = 0.0266, threshold = 0.0705 (detection 014).

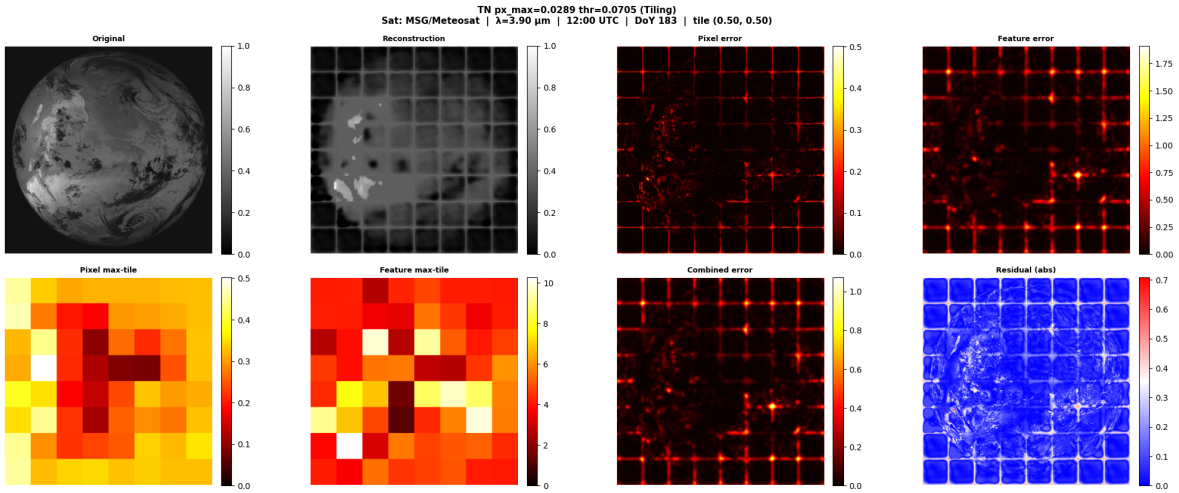


Figure C.41: True negative — px_max = 0.0289, threshold = 0.0705 (detection 003).

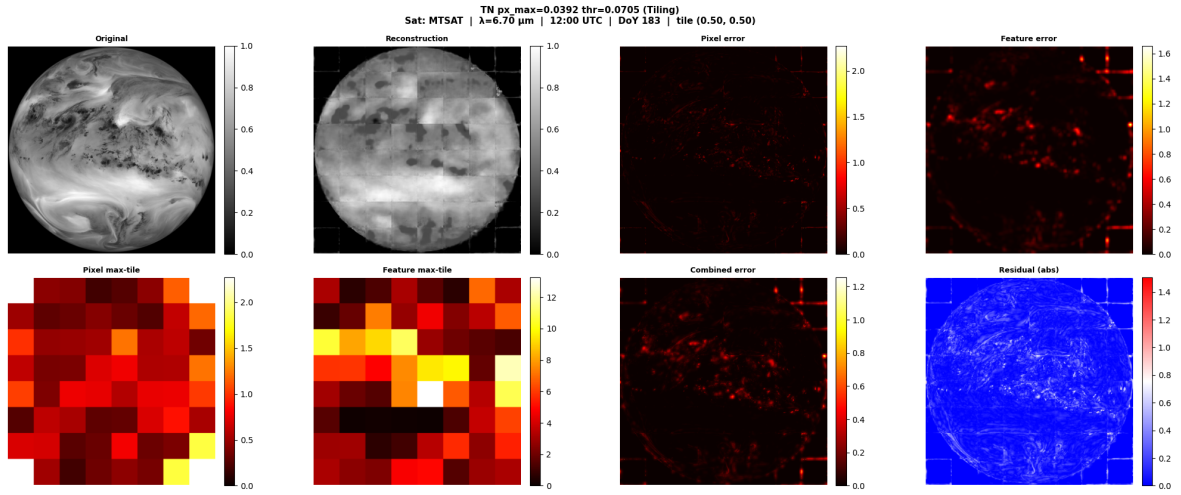


Figure C.42: True negative — px_max = 0.0392, threshold = 0.0705 (detection 023).

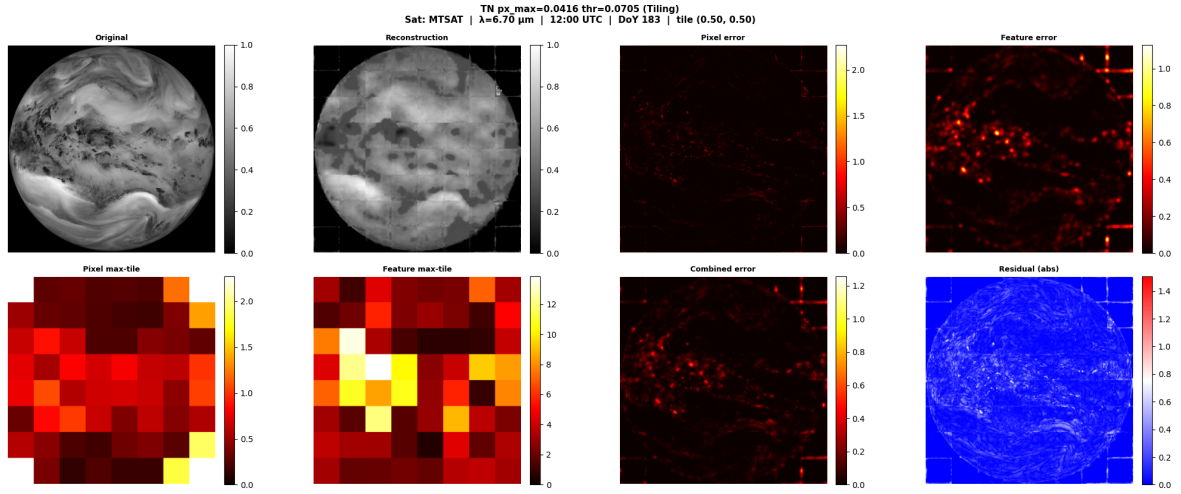


Figure C.43: True negative — px_max = 0.0416, threshold = 0.0705 (detection 024).

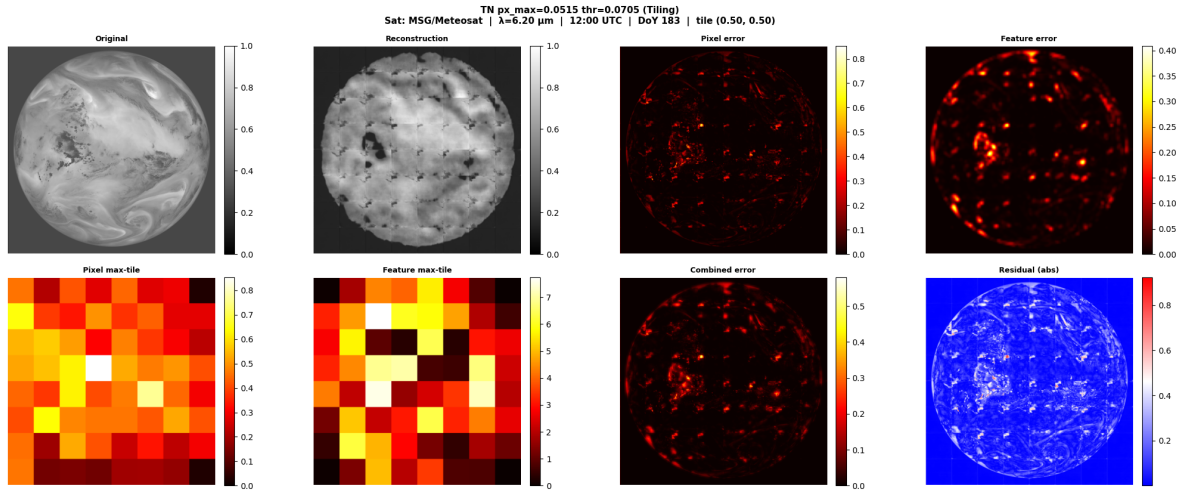


Figure C.44: True negative — px_max = 0.0515, threshold = 0.0705 (detection 015).

C.3.3. False Positives

Normal images incorrectly classified as anomalous (px_max > threshold = 0.0705).

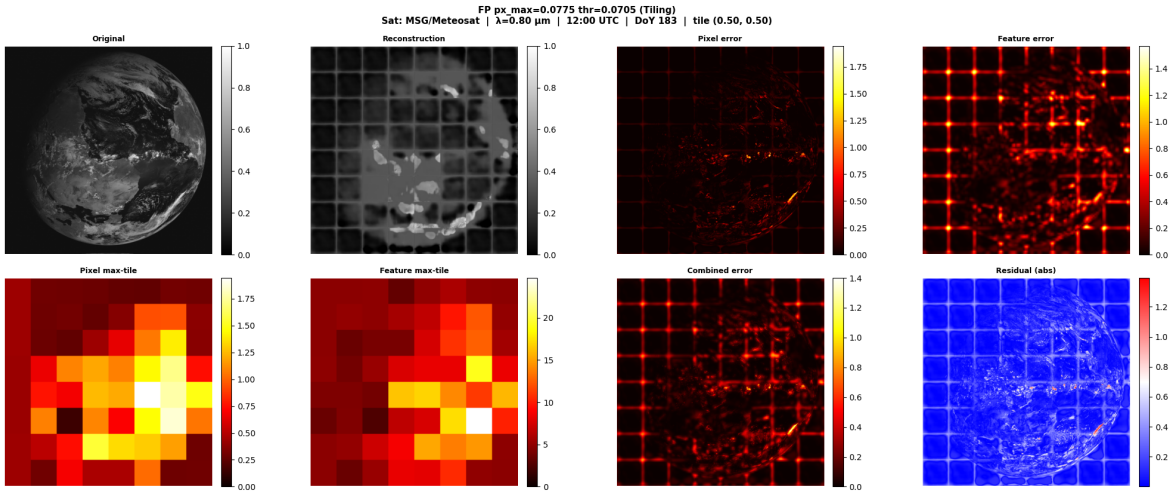


Figure C.45: False positive — px_max = 0.0775, threshold = 0.0705 (detection 013).

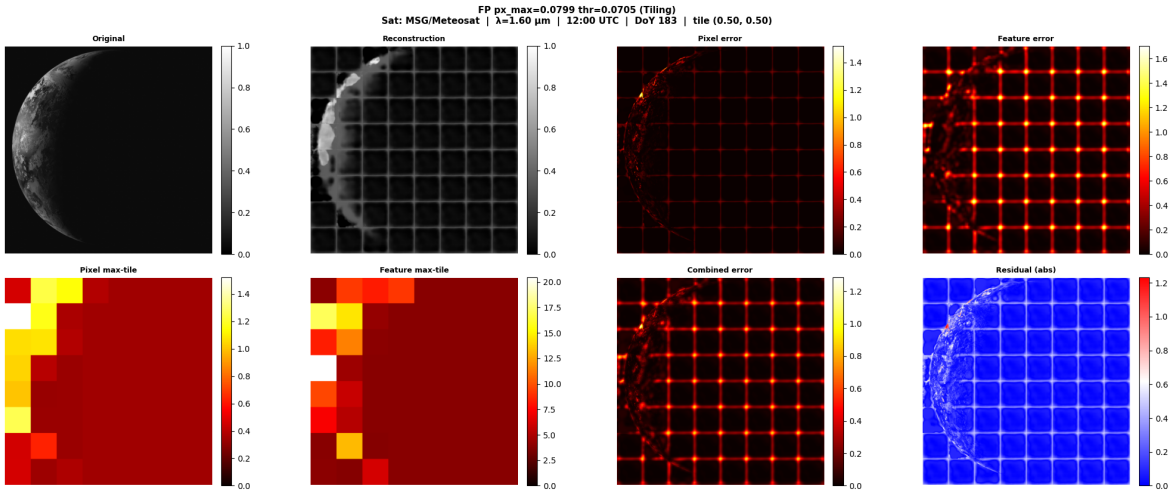


Figure C.46: False positive — px_max = 0.0799, threshold = 0.0705 (detection 001).

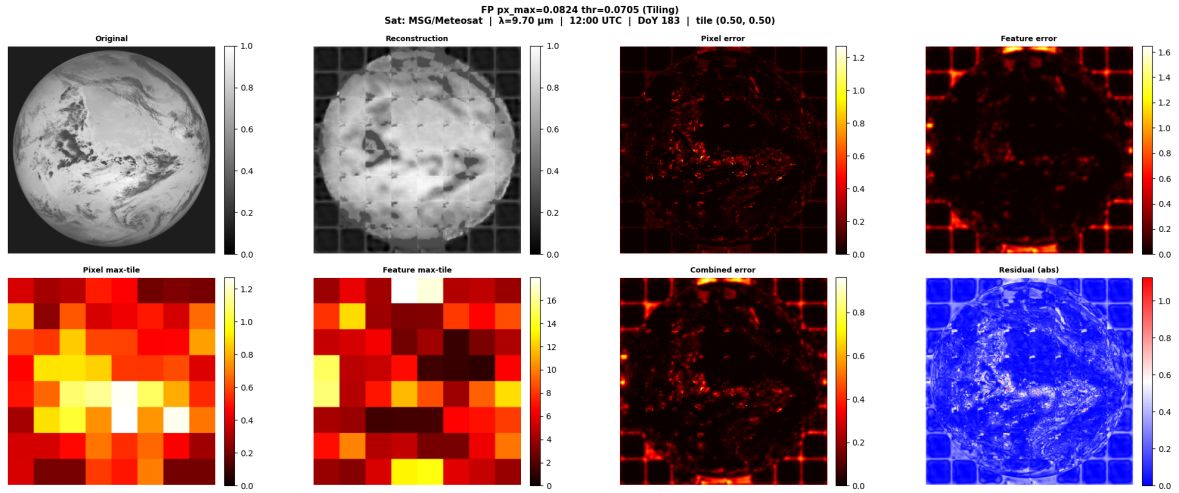


Figure C.47: False positive — px_max = 0.0824, threshold = 0.0705 (detection 005).

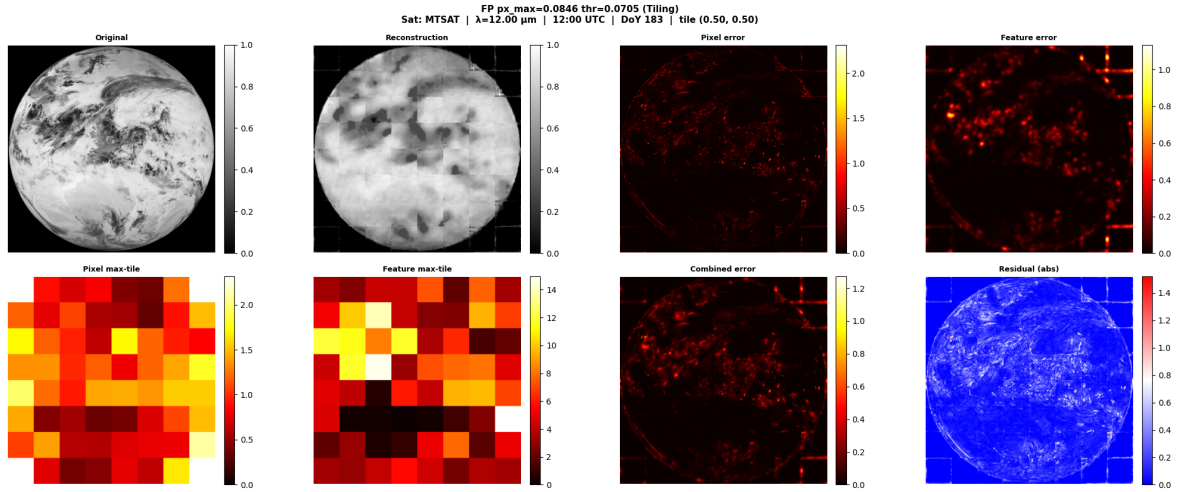


Figure C.48: False positive — px_max = 0.0846, threshold = 0.0705 (detection 020).

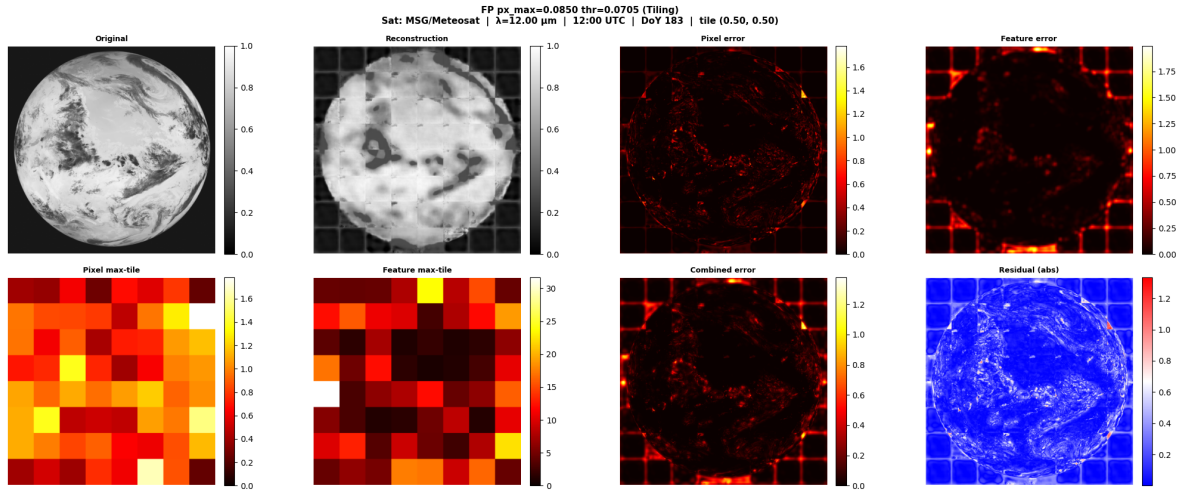


Figure C.49: False positive — px_max = 0.0850, threshold = 0.0705 (detection 009).

C.3.4. False Negatives

Anomalous images incorrectly classified as normal (px_max ≤ threshold = 0.0705). All three metadata false negatives are shown.

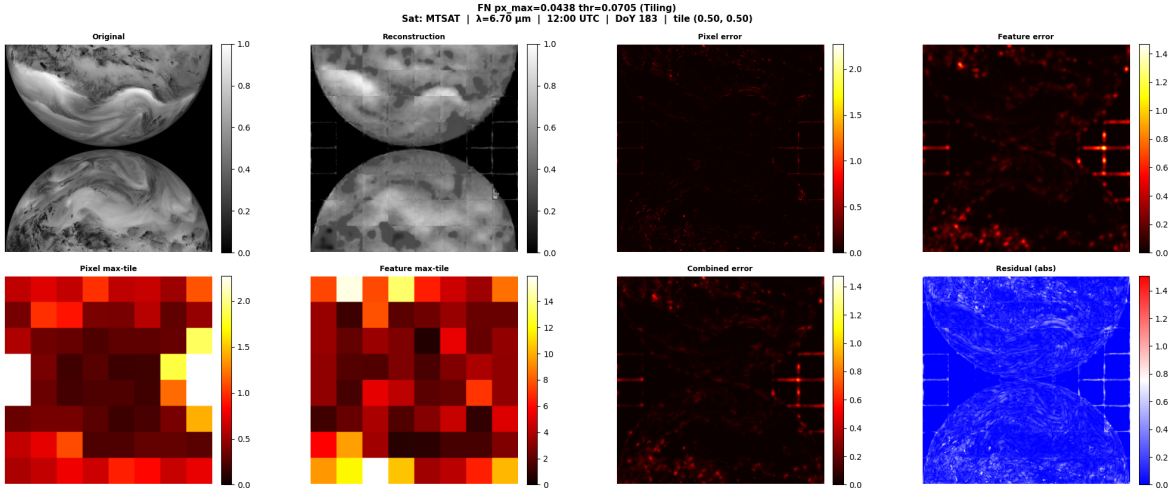


Figure C.50: False negative — px_max = 0.0438, threshold = 0.0705 (detection 023).

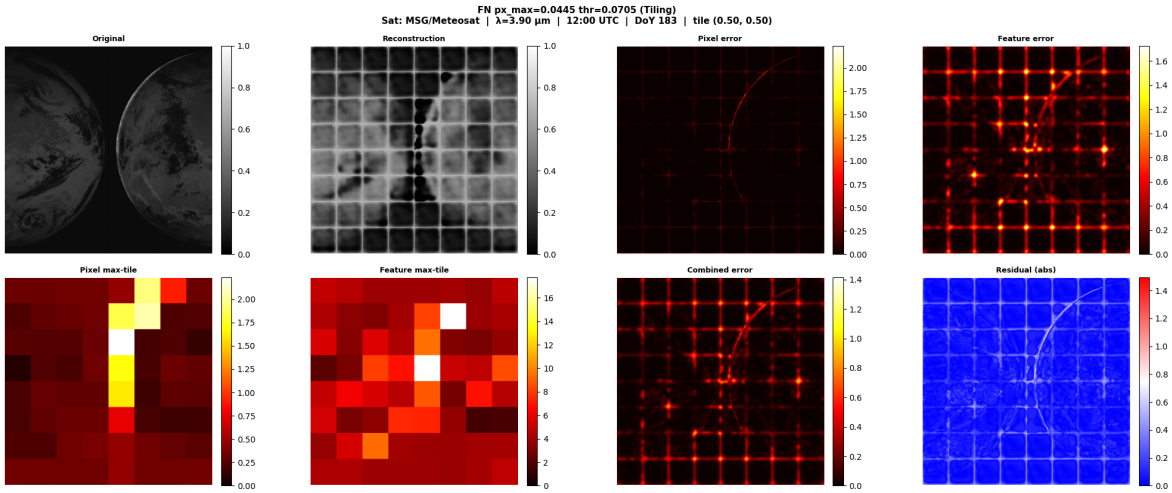


Figure C.51: False negative — px_max = 0.0445, threshold = 0.0705 (detection 002).

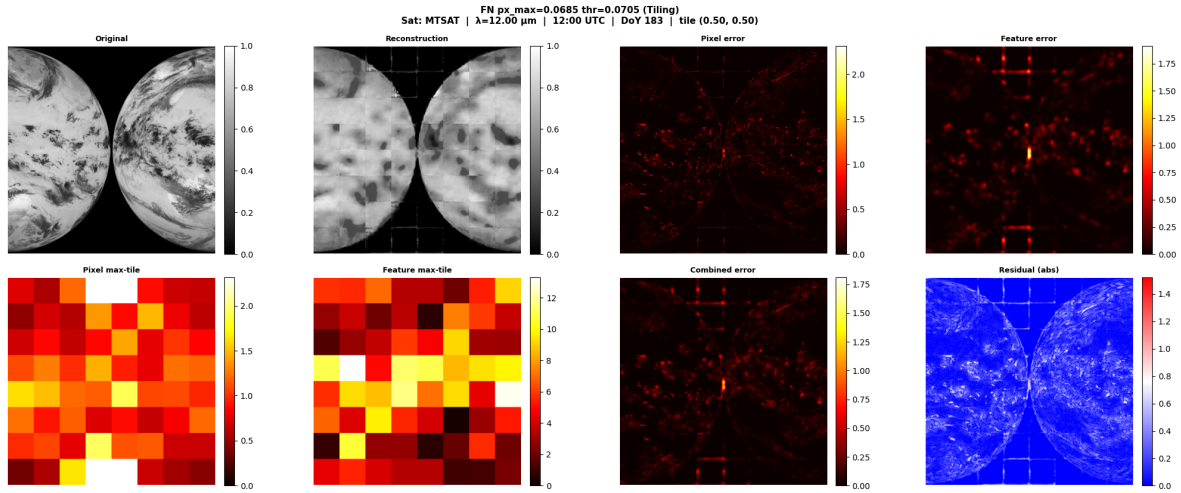


Figure C.52: False negative — px_max = 0.0685, threshold = 0.0705 (detection 020).