

Anomaly Detection for the CMS Pixel Detector Using Machine Learning

By

Tianshu Gu

B.A., Wuhan University, 2024

Thesis

Submitted in partial fulfillment of the requirements for the
Degree of Master of Science in the Department of Physics
at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2026

This thesis by Tianshu Gu is accepted in its present form
by the Department of Physics as satisfying the
thesis requirements for the degree of Master of Science

Date_____

Loukas Gouskos, Advisor

Approved by the Graduate Council

Date_____

David P. Lindstrom, Dean of the Graduate
School

Preface and Acknowledgments

This thesis presents work carried out during my Master’s program at Brown University under the supervision of Prof. Loukas Gouskos. This work builds in part upon previous work by Luka Lambrecht, including the general research direction and the associated codebase. In this thesis, I extend this framework by exploring additional methods and evaluating their performance in comparison with the previous approach.

This work would not have been possible without the support and guidance of many people.

I am sincerely grateful to Prof. Loukas Gouskos for his guidance, patience, and trust throughout this project. His insight helped shape the direction of this work, and I greatly appreciate the opportunity to learn and grow through this research.

I am especially thankful to Luka Lambrecht, whose help has been important to this project from the beginning. I benefited greatly from his previous work, technical guidance, and many thoughtful discussions.

Finally, I want to thank my family for their unconditional support and encouragement throughout my studies. I am also grateful to my friends for their companionship and for always being willing to listen to me talk about this project.

Contents

1	Introduction	1
1.1	CMS in LHC	1
1.2	DQM and Anomaly Detection	3
1.3	Research Contents and Thesis Organization	4
2	Dataset and Representation	6
2.1	Data Source and Selection	6
2.2	Pixel Detector and Occupancy Map Representation	7
3	Definition of Anomaly	9
3.1	Normal Occupancy Pattern	9
3.2	Overlapping Localized Anomalies	10
3.3	Extended Single-Layer Anomalies	11
4	VAE-Based Anomaly Detection	12
4.1	VAE Principle and Motivation	12
4.2	Detection Pipeline	14
4.2.1	Preprocessing, Reconstruction Loss, and Binarization	15
4.2.2	Single-Layer Flag	16
4.2.3	Post-processing	17
4.2.4	Combining Layers and Cross-Layer Flag	19
4.3	Results and Analysis	19

5	CLIP-Based Anomaly Detection	23
5.1	CLIP and LoRA	23
5.2	Few-shot Learning	24
5.2.1	Pipeline	24
5.2.2	Results and Analysis	26
5.3	Training with More Data via Anomaly Injection	26
5.3.1	Anomaly Injection	26
5.3.2	Results and Analysis	28
6	Conclusion	30
6.1	Summary of Results	30
6.2	Future Work	31

List of Tables

4.1	Era-wise comparison between the previous baseline and the VAE-based work-flow.	20
-----	--	----

List of Figures

1.1	The LHC collider complex.	2
1.2	The CMS detector.	3
2.1	Pixel Barrel geometry and BPix1 occupancy map.	8
3.1	Example of a normal BPix occupancy map.	10
3.2	Example of an overlapping localized anomaly in BPix1 and BPix2.	10
3.3	Example of an extended single-layer anomaly.	11
4.1	Schematic of the VAE architecture used in this work.	13
4.2	Overall pipeline of the VAE-based anomaly-detection method.	14
4.3	Example of preprocessing, reconstruction, loss calculation, and binarization.	15
4.4	Example of the Single-Layer Flag.	16
4.5	Example of per-layer post-processing from binary loss to rebinned loss.	18
4.6	Example of layer combination and the Cross-Layer Flag.	19
4.7	Example of an anomaly identified by the VAE-based workflow but missed by the previous baseline.	21
5.1	Preprocessing for CLIP few-shot learning.	25
5.2	Example of anomaly injection on Era C.	27

Chapter 1

Introduction

1.1 CMS in LHC

The Standard Model of particle physics provides the theoretical framework for describing the known fundamental particles and their interactions. Experimental tests of this framework, as well as searches for phenomena beyond it, require collisions at extremely high energies, very large collision rates, and detectors capable of recording the resulting particle signatures with high precision. In this context, the Large Hadron Collider (LHC) provides one of the most important experimental platforms in contemporary high-energy physics.

The LHC is the world's largest and most powerful particle accelerator [1]. It consists of a 27-kilometer ring equipped with superconducting magnets and accelerating structures that guide and accelerate particle beams to near-light speed before bringing them into collision at several interaction points around the ring. It is the final stage of the CERN accelerator complex, shown schematically in Figure 1.1. Four major experiments are located at these collision points: ATLAS, CMS, ALICE, and LHCb. The combination of high collision energy and intensity makes it possible to study rare processes and precision observables that are central to modern particle physics research.

CMS is one of the two large general-purpose experiments at the LHC. A schematic overview of the detector is shown in Figure 1.2. The detector has a cylindrical geometry and is organized in several concentric layers around the interaction point, in a configuration often

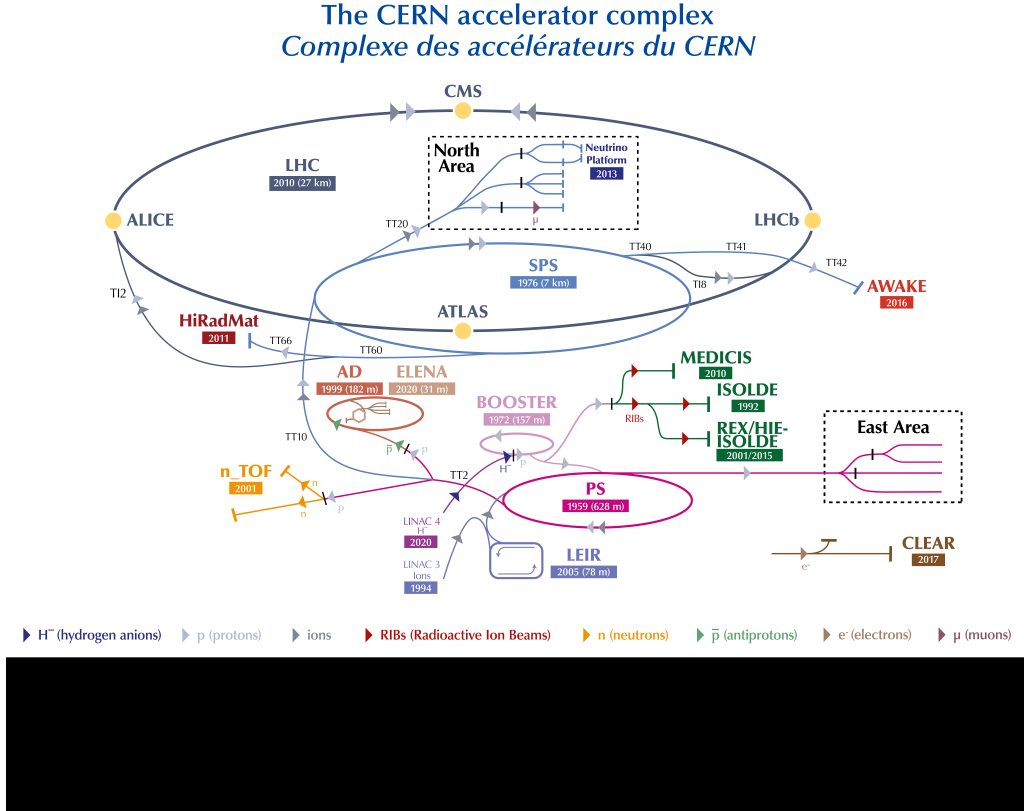


Figure 1.1: The LHC collider complex [1].

described as onion-like. From the inside outward, the main subsystems are the silicon tracker, the electromagnetic calorimeter, the hadronic calorimeter, the superconducting solenoid, and the muon system. These subdetectors work together to reconstruct charged-particle trajectories, measure electromagnetic and hadronic energy deposits, and identify muons. This integrated design allows CMS to record collision events with high precision and supports a broad physics programme, from precision measurements within the Standard Model to searches for new particles and interactions [2].

Within CMS, this thesis focuses on the pixel detector, which forms the innermost part of the silicon tracking system. Because it is located closest to the beam pipe, the pixel detector provides the earliest and most precise spatial measurements of charged particles emerging from the collision point. These measurements are essential for reconstructing the tracks of charged particles produced in the collision, as well as for determining the primary vertex. In

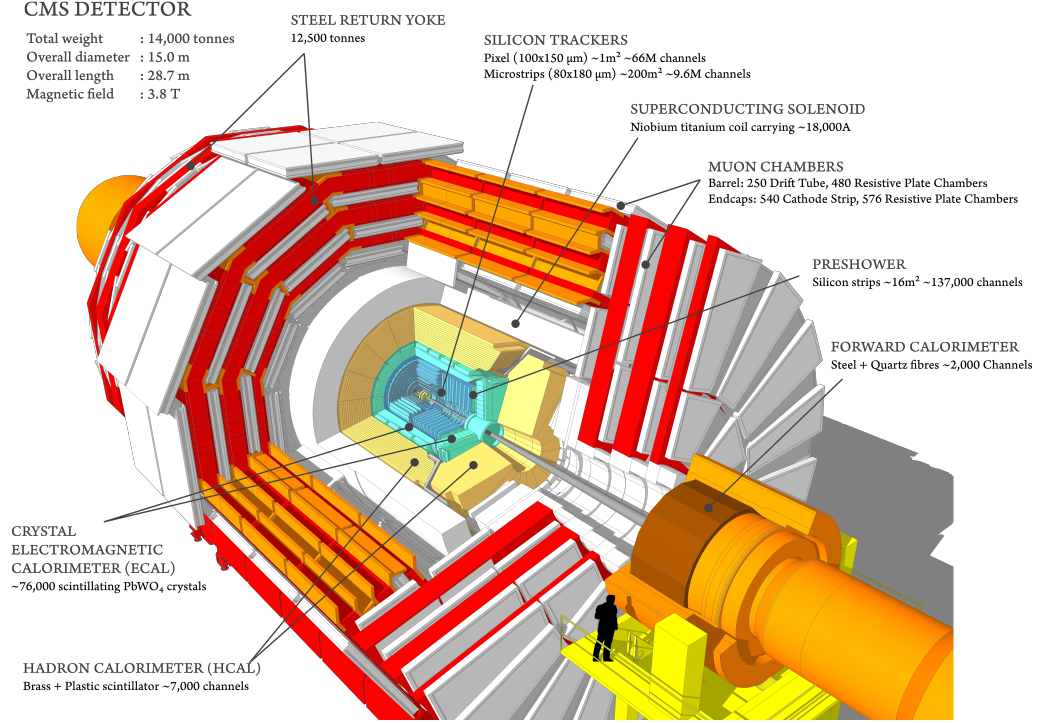


Figure 1.2: The CMS detector [2].

the Barrel Pixel (BPix) region considered in this work, the detector is organized into four cylindrical layers. The highly structured geometry of these layers is directly reflected in the monitoring images used later and constitutes one of their most important features.

1.2 DQM and Anomaly Detection

Given the complexity of the CMS detector and the large volume of data produced during operation, a dedicated Data Quality Monitoring (DQM) workflow is essential for ensuring stable detector performance and reliable physics data. In CMS, DQM serves monitoring, certification, and debugging functions throughout the data-taking and reconstruction chain [3].

In the current DQM procedure, detector information is typically processed into monitoring elements, such as histograms and related summary views, which are inspected by shifters in real time and later reviewed by subsystem experts. These checks are used both to identify

detector or readout problems and to determine whether the recorded data should be certified as suitable for physics analysis. In practice, data quality is still often judged from run-level aggregated views spanning long periods of data taking, from many minutes to several hours. This procedure has several well-known limitations [4]: the number of histograms to inspect is large, the outcome depends strongly on how the underlying information is visualized, and long monitoring shifts can lead to fatigue, oversights, and subjective judgments. Despite ongoing modernization efforts, the overall workflow remains largely human-centered.

An important recent development is the deployment of Data Inspector for Anomalous Luminosity Sections (DIALS) in 2024, a platform that stores and displays histograms at lumisection (LS) level rather than only at run level [5]. Since one lumisection corresponds to about 23 seconds of data taking, DIALS significantly improves the temporal resolution of monitoring and makes it possible to resolve short, localized anomalies that would previously be hidden inside longer runs. At the same time, this finer granularity greatly increases the number of units that must be inspected, making purely manual inspection less scalable and less reliable. DIALS therefore makes short anomalies visible, but also strengthens the need for automated anomaly detection at LS level, not to replace experts, but to guide their attention toward suspicious cases more efficiently and support more reliable data certification.

1.3 Research Contents and Thesis Organization

Research on detector anomaly detection in CMS has already explored several methodological directions, including script-based or rule-based procedures, matrix-factorization-based approaches such as Non-negative Matrix Factorization (NMF) [6], and reconstruction-based methods such as autoencoders [7]. These approaches each offer particular advantages. Rule-based methods are usually easy to interpret and closely aligned with detector expertise, but they often require substantial manual tuning and can be difficult to generalize. Methods such as NMF provide more data-driven decompositions while retaining some interpretability, but their modeling capacity is limited when the underlying patterns become more complex.

Autoencoder-based approaches are more flexible and have shown promising performance in detecting subtle anomalous structures, yet they still face challenges in turning reconstruction errors into robust final decisions and often depend on additional post-processing or detector-specific heuristics.

Motivated by both the strengths and the limitations of these earlier approaches, this thesis builds on prior work while exploring two machine-learning directions. The first direction is based on a variational autoencoder (VAE), which extends the reconstruction-based paradigm to model multi-layer pixel occupancy images. In this part of the study, reconstruction errors are converted into lumisection-level anomaly decisions through a pipeline that follows the detector geometry and the anomaly definitions adopted in this thesis. The second direction asks whether more recent representation-learning methods can provide a complementary and more direct alternative. For this purpose, a CLIP-based visual encoder is studied together with Low-Rank Adaptation (LoRA) fine-tuning, and an anomaly-injection scheme is introduced to compensate for the limited number of real anomalous samples. Rather than assuming that one of these approaches is universally superior, this thesis treats them as two stages in an evolving research process.

The remainder of this thesis is organized as follows. Chapter 2 describes the dataset and its representation used in this work, including the form of the monitoring images and the main characteristics of the data under study. Chapter 3 defines the anomaly definitions considered in the analysis. Chapter 4 presents the VAE-based method, including the model principle, the full anomaly-detection pipeline, and the corresponding results. Chapter 5 presents the CLIP-based method and the anomaly-injection algorithm. Chapter 6 summarizes the results and discusses future work.

Chapter 2

Dataset and Representation

2.1 Data Source and Selection

The dataset used in this work is derived from the CMS DQM system and covers the 2024 data-taking period. It is based on a collection of collision events, selected essentially at random to avoid bias and processed with a first pass of the full event reconstruction algorithm. The monitoring outputs are organized in the detector mechanical coordinate system and stored at lumisection level, so that each sample corresponds to a short interval of detector operation, about 23 seconds.

Within the CMS data-taking hierarchy, the LS is the finest unit considered in this work. Consecutive LS form a run, and runs are grouped into broader eras associated with common operational or reconstruction conditions. In the 2024 dataset, the eras range from A to J. Since A and J are the earliest and latest eras and contain substantial commissioning or debugging periods, only eras B to I are included in this study, corresponding to a total of 564,275 lumisections.

Lumisections are selected based on some automatic operational flags that require the accelerator to be in stable-beams conditions and the tracking detector to be active. Additional event-quality requirements are then applied: each BPix layer must have a minimum total occupancy larger than $0.5\text{M}/(\text{layer number})$, the pileup must satisfy > 25 , and the ZeroBias trigger rate must satisfy $> 10\text{ Hz}$. Across the 2024 eras from B to I considered in this thesis,

the Detector Control System (DCS)-related filters remove 254,219 lumisections, while the event-quality filters remove 271,155. Because some lumisections fail both sets of criteria, the combined preselection removes 271,997 lumisections in total, corresponding to about 48.2% of the full sample. The retained dataset therefore reflects detector behavior under operationally meaningful conditions rather than trivial differences caused by low statistics or non-standard data-taking periods.

2.2 Pixel Detector and Occupancy Map Representation

Within the selected dataset, the analysis focuses on the Barrel Pixel (BPix) region of the CMS pixel detector, which is organized into four cylindrical layers around the beam pipe, denoted as BPix1 to BPix4. For each LS, the data are represented by one two-dimensional occupancy map per layer. These maps are defined at the level of the Readout Chip (ROC).

Each ROC consists of a matrix of 80×52 pixels, which form the fundamental detection elements recording individual hits. Adjacent pixel hits produced by a single particle crossing are grouped into clusters, which provide a higher-level representation of detector activity. For a given lumisection, each bin in the occupancy map records the number of clusters associated with the corresponding ROC during that time interval. The occupancy value can therefore be interpreted as an approximate measure of the local particle flux.

Figure 2.1 illustrates how detector geometry is mapped to image representation. In the left panel, the axial cross-section of the Pixel Barrel detector shows the ladder structure of the barrel layers, where each black bar corresponds to a ladder and the ladders are segmented into modules along the beam direction. In the right panel, the BPix1 occupancy map shows the corresponding ROC-level image representation. At this scale, one detector module corresponds to a 2×8 block of ROCs. This is important because many detector effects appear at the module scale, rather than at isolated single-bin level. Across the four BPix layers, the maps share the same layout in the module direction, corresponding to eight modules along the beam direction, while the number of ladders increases with detector radius: 12, 28, 44,

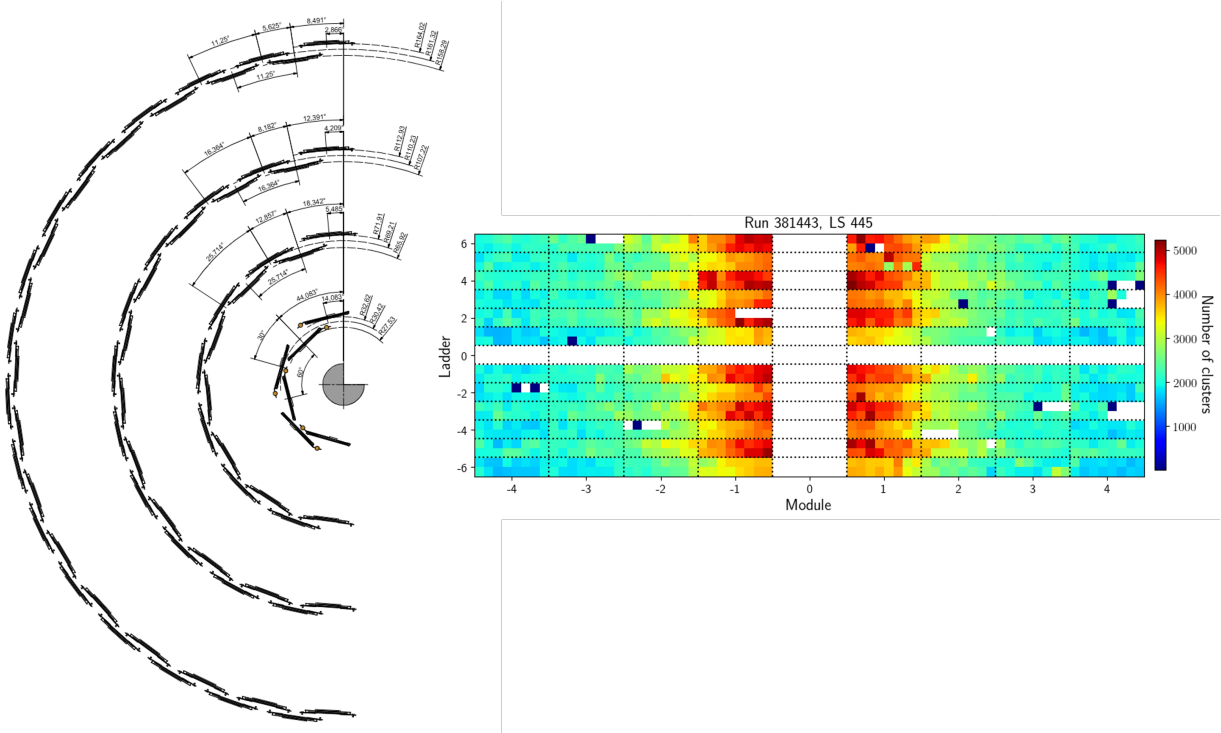


Figure 2.1: Left: axial cross-section of the Pixel Barrel detector [8]. Right: example of a BPix1 occupancy map.

and 64 for BPix1 to BPix4, respectively.

This representation defines the data domain used in the remainder of the thesis. Because the maps encode detector structure explicitly, spatial patterns in the occupancy maps can be interpreted directly in terms of the underlying detector geometry.

Chapter 3

Definition of Anomaly

In this thesis, defining what should count as an anomaly is not entirely straightforward. In almost every lumisection, a small number of individual ROCs or small groups of ROCs may be in an error state or may temporarily record little or no occupancy. Such effects can arise, for example, from short-lived problems in powering or readout electronics, potentially related to the extreme radiation environment, although they are often recovered automatically after about one minute. They therefore do not by themselves imply that the lumisection should be classified as anomalous. Taking these considerations into account, the anomalies of interest in this thesis are divided into two main categories: localized anomalous regions that overlap in geometrically corresponding positions across two or more layers, and extended anomalies that affect a large fraction of a single layer.

3.1 Normal Occupancy Pattern

Under normal operating conditions, the occupancy map exhibits a regular pattern shaped by detector geometry and collision topology. As illustrated in Figure 3.1, the cluster occupancy is generally higher near the central module region and decreases toward both sides of the map. This trend is consistent with the fact that charged particles originate from collisions near the interaction point and intersect the barrel layers in a geometrically non-uniform way. Apart from a small number of isolated low-occupancy ROC regions, the overall structure remains smooth and consistent with the expected detector response.

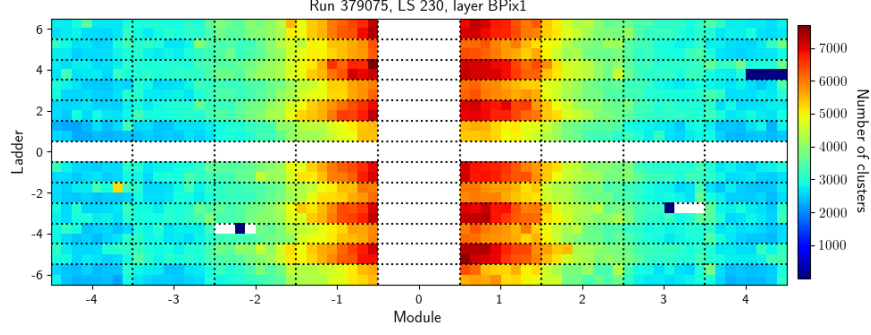


Figure 3.1: Example of a normal BPix occupancy map.

3.2 Overlapping Localized Anomalies

Figure 3.2 shows a representative example of an overlapping localized anomaly using two layers from the same lumisection. Although the overall occupancy pattern in each map remains broadly similar to that of a normal sample, the red boxes mark anomalous low-occupancy regions that appear in geometrically corresponding positions. This situation is treated as anomalous because correlated cluster deficits across layers can degrade track reconstruction and therefore affect downstream physics analysis. By contrast, a similar localized deviation confined to a single layer is not treated here as anomalous, because missing information in one layer can often still be recovered by the track reconstruction algorithm.

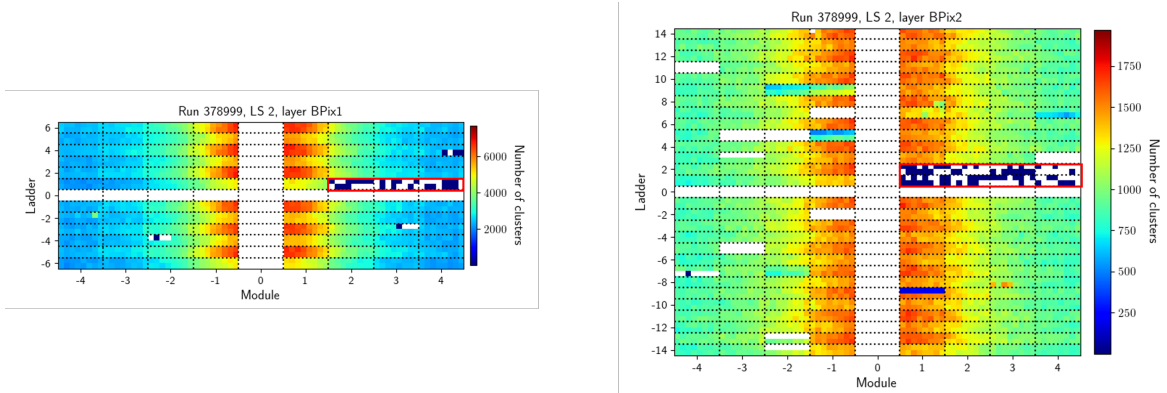


Figure 3.2: Example of an overlapping localized anomaly in BPix1 and BPix2.

3.3 Extended Single-Layer Anomalies

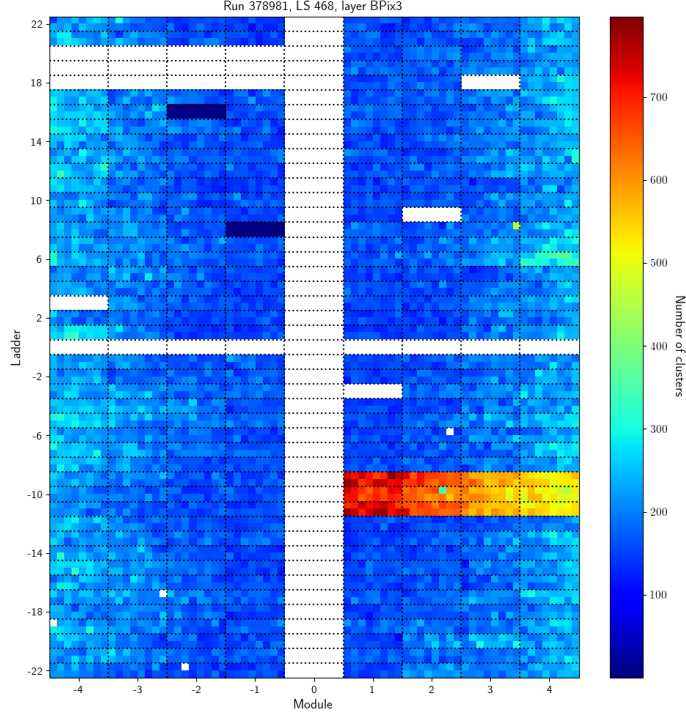


Figure 3.3: Example of an extended single-layer anomaly.

Figure 3.3 shows a representative extended anomaly confined to a single layer. Unlike the localized case, the anomalous behavior is not restricted to one compact region. Instead, most of the layer departs from the expected occupancy pattern, with only a limited subset of modules remaining close to normal. In the example shown here, the map displays an unusual large-scale structure in which the occupancy is comparatively higher on both sides and lower in the central region, in contrast to the normal pattern introduced in Figure 3.1. This type of anomaly is treated as significant because the number of affected ROCs is large and the deviation is clearly incompatible with ordinary detector behavior, which can be the indicator of malfunctioning pixel modules.

Chapter 4

VAE-Based Anomaly Detection

As discussed in Chapter 1, earlier CMS anomaly-detection studies have already shown that reconstruction-based methods are well suited to DQM images, because the task is dominated by normal detector behavior while truly anomalous lumisections remain comparatively rare. The VAE-based method adopted in this work follows the same general idea, but extends it in a way that is better matched to the structured four-layer BPix data described in Chapter 2 and to the anomaly definitions introduced in Chapter 3.

4.1 VAE Principle and Motivation

A variational autoencoder (VAE) is a generative model that learns a low-dimensional latent representation of the input data and reconstructs the input from this latent space [9]. In contrast to a plain autoencoder, the encoder does not map an input directly to one deterministic latent vector. Instead, it predicts the parameters of a latent distribution, from which a latent vector is sampled and then passed to the decoder. The training objective combines reconstruction quality with a regularization term in latent space, so that the learned representation does not simply memorize the training samples but the distributions.

This model is well suited to anomaly detection when the dataset is dominated by normal samples. Under this condition, the VAE mainly learns the regular spatial organization of normal cluster occupancy maps. A normal lumisection can therefore be reconstructed with relatively small error, whereas an anomalous lumisection tends to produce a larger mismatch

between the input and the reconstruction. In this sense, the method is unsupervised and fully data-driven: it does not require a labeled training set and does not depend on predefined assumptions about specific anomaly types.

The choice of a VAE in this work is motivated by both the data characteristics and the limitations of earlier approaches discussed in Chapter 1. In CMS pixel DQM, anomalous lumisections are comparatively rare, which makes reconstruction-based methods a natural starting point. Compared with script-based monitoring, the VAE reduces the need to enumerate detector rules in advance for every possible anomalous shape. Compared with matrix-factorization methods such as NMF, the nonlinear encoder-decoder structure provides a more flexible description of normal occupancy patterns. Compared with a plain autoencoder, the variational formulation introduces a more constrained latent representation and reduces the risk of overfitting rare irregular structures.

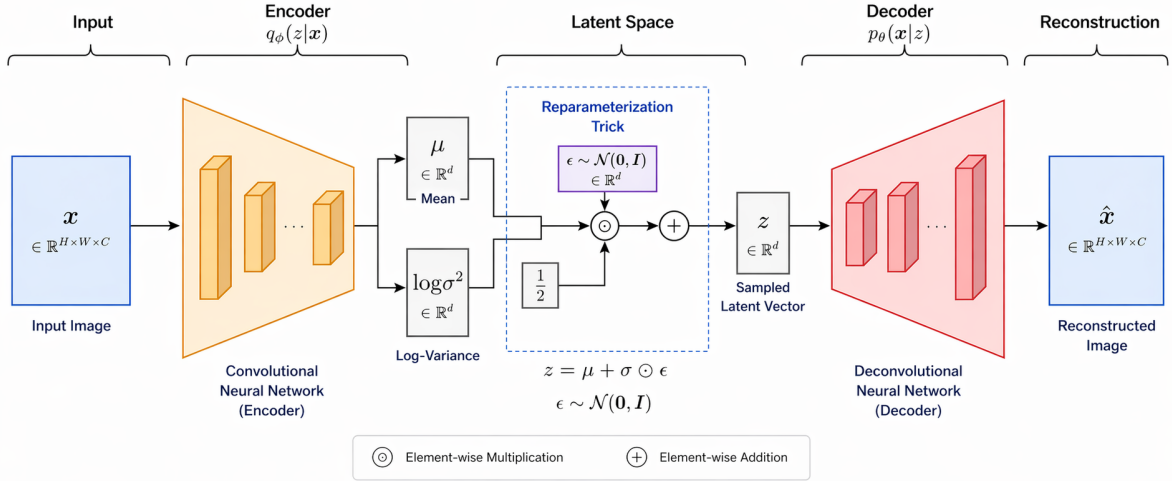


Figure 4.1: Schematic of the VAE architecture used in this work.

Figure 4.1 illustrates this encoder-latent-decoder structure. The input image is first mapped by the encoder to the parameters of a latent distribution, represented by the mean and log-variance. A latent vector is then sampled through the reparameterization step and passed to the decoder to produce a reconstruction of the input. The training objective contains two terms: a reconstruction term, which encourages the output to remain close to the

input, and a Kullback–Leibler regularization term, which constrains the latent distribution toward a standard normal prior. In compact form, the loss can be written as

$$L = L_{\text{rec}} + \beta L_{\text{KL}}.$$

4.2 Detection Pipeline

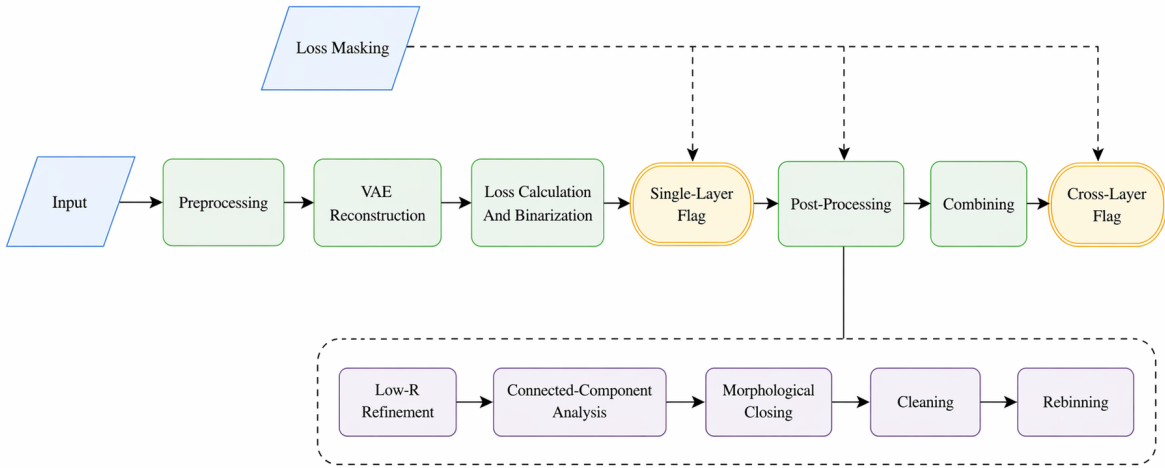


Figure 4.2: Overall pipeline of the VAE-based anomaly-detection method.

Figure 4.2 gives the complete pipeline of the VAE-based method. Starting from the input occupancy maps, the workflow first proceeds through preprocessing, VAE reconstruction, loss calculation, and binarization for each BPix layer separately. After these steps, a single-layer flag is applied to determine whether the lumisection should be classified as an extended single-layer anomaly, corresponding to the anomaly type defined in Section 3.3. Lumisections that are not flagged at this stage continue to the subsequent post-processing steps.

After the single-layer decision, the remaining loss maps undergo per-layer post-processing and are then combined across layers. A final cross-layer flag is applied to the combined map in order to identify overlapping localized anomalies, as defined in Section 3.2. Throughout this workflow, a long-term loss mask is used to suppress known persistently inactive regions

so that they do not dominate the later anomaly decisions.

4.2.1 Preprocessing, Reconstruction Loss, and Binarization

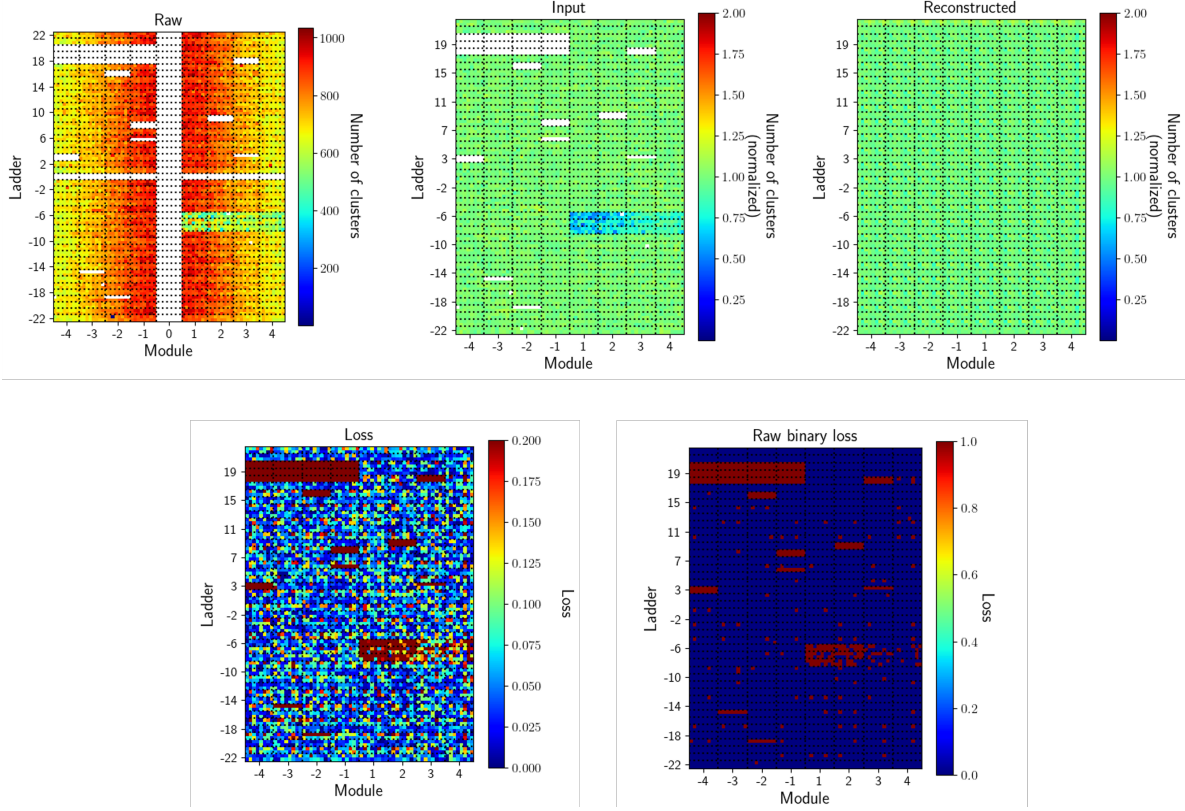


Figure 4.3: Example of preprocessing, reconstruction, loss calculation, and binarization.

At this stage, each BPix layer is processed independently. For each layer, the empty dummy bins are first removed from the raw occupancy map. The cropped map is then divided bin by bin by the average occupancy map from the previous era. This choice is motivated by the intended online application, where the full data of the current era are not yet available when one wants to normalize a lumisection of this era. It is also motivated by the need to suppress the strong built-in non-uniformity of the raw occupancy pattern, so that the model focuses more on deviations from the expected structure. The normalized map is further rescaled so that its average value is one, which makes the input more suitable for VAE training and inference. The VAE used in this study is trained on lumisections from Era

C. Although Era A is excluded from the dataset, Era B is still considered too early and not yet sufficiently stable for training, so Era C is taken as the earliest stable era suitable for this purpose. It contains 24,353 lumisections after filtering, corresponding to about 8.3% of the retained dataset. During training and inference, bins with value 0 are replaced by 1, which corresponds to the expected average value after preprocessing, in order to avoid making frequently occurring empty regions a dominant feature for the model to learn. The trained model then reconstructs the preprocessed input, the absolute difference between input and reconstruction is taken as the loss map, and thresholding this loss yields the raw binary loss map that is used in the following stages.

Figure 4.3 shows one representative per-layer example of this sequence. The upper-left panel is the raw occupancy map. The upper-middle panel is the corresponding input after cropping and normalization, and the upper-right panel is the reconstruction produced by the trained VAE. The lower-left panel shows the loss map obtained from the absolute difference between the input and the reconstruction, while the lower-right panel shows the raw binary loss map after thresholding. In this example, the localized low-occupancy region is visible in the input, remains prominent in the loss map, and is retained in the binary loss map together with some scattered high-loss pixels.

4.2.2 Single-Layer Flag

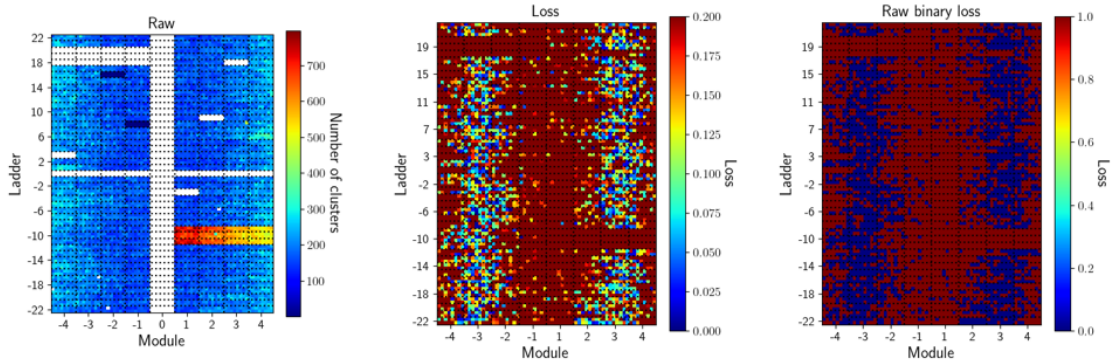


Figure 4.4: Example of the Single-Layer Flag.

The first anomaly decision is the Single-Layer Flag. For each layer, the ratio

$$R = \frac{N_{\text{flagged}} - N_{\text{masking}}}{N_{\text{total}}}$$

is evaluated on the raw binary loss map. If R exceeds the predefined threshold in any layer, the entire lumisection is directly flagged. This rule is intended to identify extended single-layer anomalies, for which the anomalous region is already sufficiently large at layer level.

Figure 4.4 shows such a case. In the raw occupancy map, the deviation is not confined to one compact region but extends over a large fraction of the layer. The same behavior is reflected in the loss map and in the raw binary loss map, where flagged pixels occupy a substantial part of the image. For this reason, the lumisection is directly classified by the Single-Layer Flag as an extended single-layer anomaly, without requiring the later cross-layer steps.

By contrast, the example in Figure 4.3 contains a localized low-occupancy region, but the fraction of flagged pixels remains below the Single-Layer Flag threshold. That case is therefore not rejected at this stage and is passed to the subsequent post-processing and cross-layer analysis.

4.2.3 Post-processing

The purpose of per-layer post-processing is to remove small and less important anomalous regions from the raw binary loss map while preserving and repairing the larger structures that are more likely to require attention.

Figure 4.5 shows this procedure in detail. Starting from the raw binary loss map, a Low-R refinement, which means the top fraction of loss pixels is retained to form the binary loss map, is first applied when the abnormal-pixel ratio is small. This preserves the most informative part of a weak anomaly instead of letting it disappear under a fixed threshold.

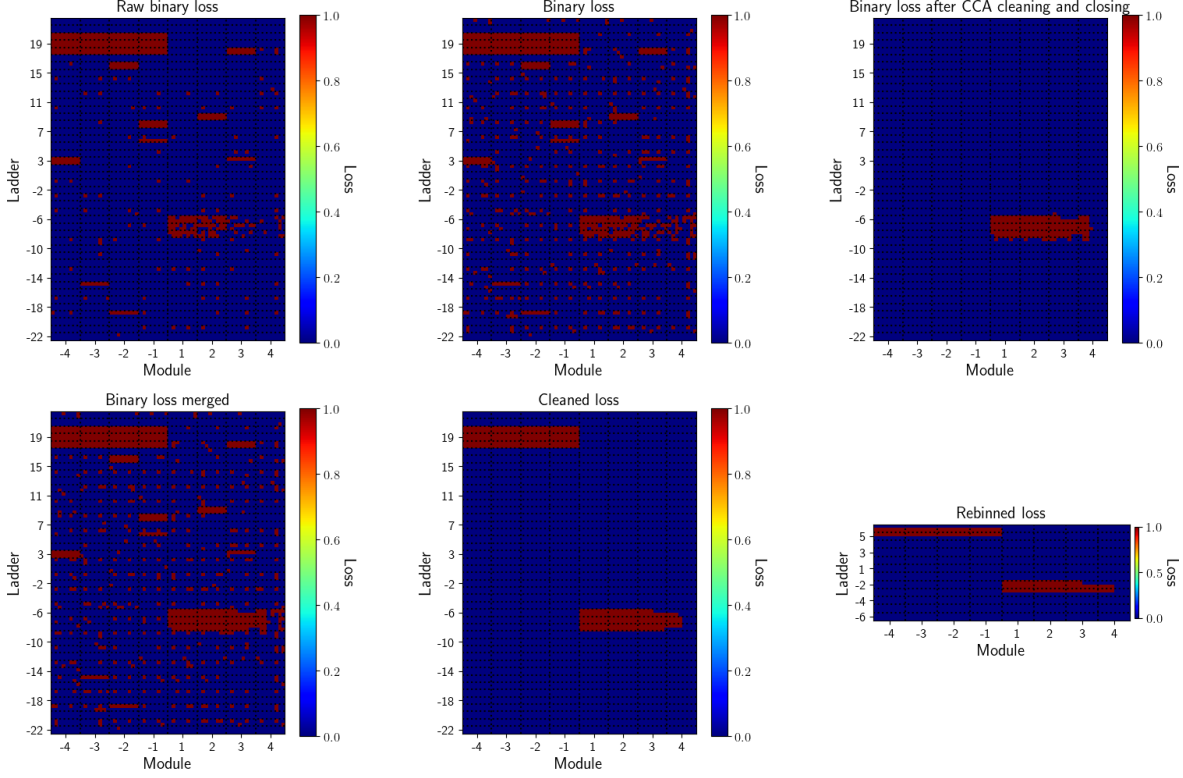


Figure 4.5: Example of per-layer post-processing from binary loss to rebinned loss.

Connected-component analysis is then applied to isolate sufficiently large clustered high-loss regions before the closing operation, so that unrelated small anomalies are not incorrectly merged during morphological closing. Closing is used to fill small holes inside the retained region, producing the compact structure shown in the upper-right panel. This result is then merged with the original binary loss map to restore smaller structures that were removed by the connected-component step. The following cleaning step keeps only clusters of size 2×16 or larger, that is, at least two full neighboring modules, while allowing one ROC with lower loss. Finally, each layer map is rebinned to the BPix1 shape, which is used in the next step for combining. As shown in Figure 4.5, this sequence removes many scattered high-loss pixels while preserving the larger elongated anomalous structures.

4.2.4 Combining Layers and Cross-Layer Flag

After per-layer processing, the rebinned loss maps from different layers are combined by a simple sum.

The Cross-Layer Flag is then applied to the combined map. In the configuration used here, the lumisection is flagged if the combined loss map contains a high-loss cluster of at least size 1×8 or 2×4 . This step corresponds to the anomaly definition in Chapter 3, where a localized anomaly becomes significant only when it is supported by multiple layers rather than appearing in isolation in a single map.

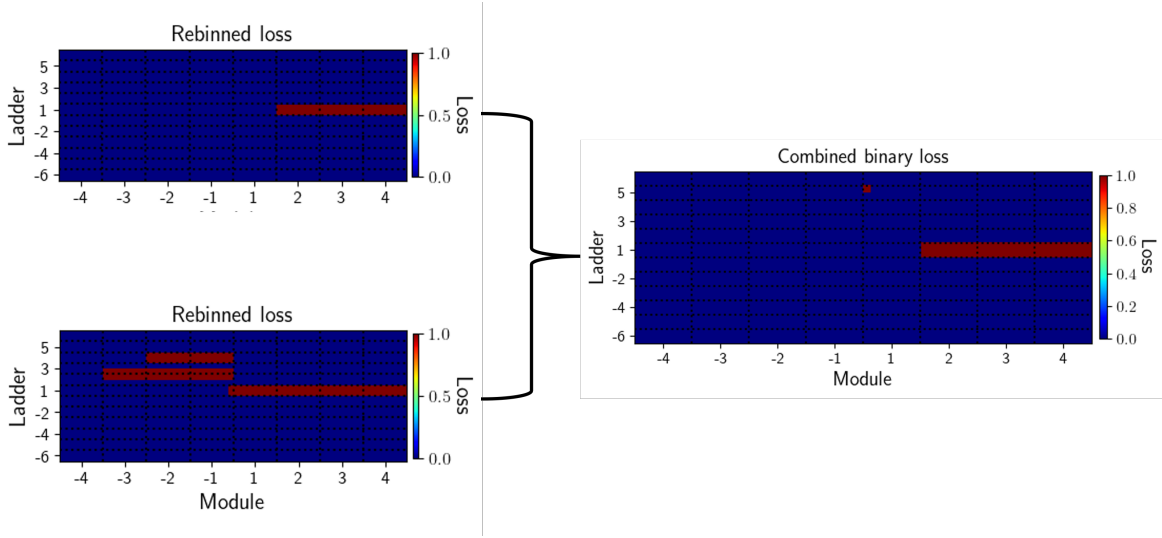


Figure 4.6: Example of layer combination and the Cross-Layer Flag.

Figure 4.6 shows a representative example. The two left panels display rebinned loss maps from two different layers, and the right panel shows the combined binary loss map obtained from their sum. In this example, the horizontal high-loss region appears in corresponding positions in both layers and is therefore retained clearly in the combined map.

4.3 Results and Analysis

Table 4.1 summarizes the era-wise detection results of the VAE-based workflow in comparison with the previous NMF-based baseline [10]. Overall, the new results remain consis-

Era	Previous	New	Lost	Now
B-v1	212	48	0	260
C-v1	48	1	1	48
D-v1	18	10	0	28
E-v1	17	10	0	27
E-v2	10	1	0	11
F-v1	95	19	0	114
G-v1	262	4	0	266
H-v1	27	9	2	34
I-v1	8	0	1	7
I-v2	55	32	0	87

Table 4.1: Era-wise comparison between the previous baseline and the VAE-based workflow. All the lost cases are due to different filtering conditions.

tent with the previous ones. Almost all lumisections flagged by the previous method remain flagged here and, at the same time, the VAE-based workflow identifies additional anomalous lumisections in several eras.

Figure 4.7 shows an example of an anomaly that is identified by the present VAE-based workflow but not by the previous baseline. It consists of overlapping localized anomalous regions in layers 2 and 4, marked by red boxes. In layer 2, the anomalous region has zero occupancy. In layer 4, the anomalous region has occupancy that is lower than the normal local level but not fully zero, so the deviation is more subtle and harder to detect.

It is also worth noting that a substantial fraction of the additional flagged lumisections correspond to extensions of intervals that were already identified as anomalous by the previous method. In this sense, the VAE-based method preserves the main behavior of the earlier approach while improving sensitivity to different anomalies. This is driven by the combined effects of modifications to the model and the post-processing.

One important strength of this method is that it preserves a high degree of interpretability. Each intermediate output in the pipeline has a clear meaning: the reconstructed image represents the learned normal expectation, the loss map shows the deviation from that expectation, the binary loss map marks candidate anomalous ROCs, and the combined map

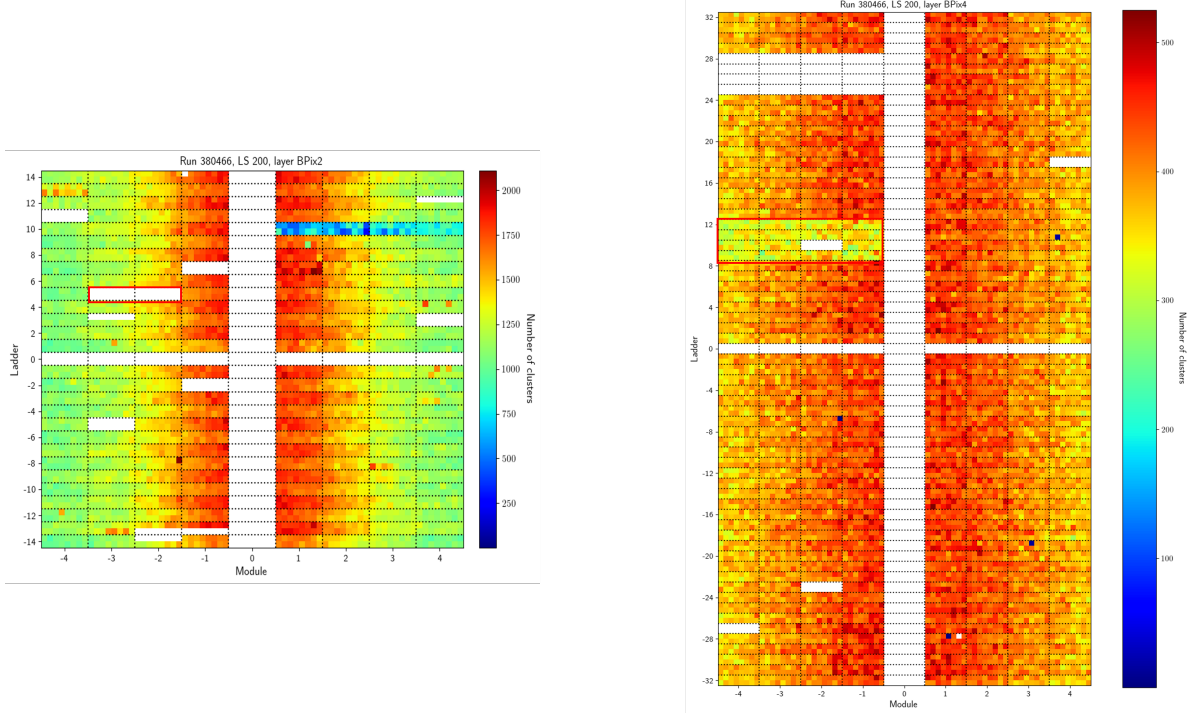


Figure 4.7: Example of an anomaly identified by the VAE-based workflow but missed by the previous baseline.

shows where anomaly evidence is supported across layers. This makes it possible to trace a flagged lumisection back to concrete image structures rather than to a single opaque score. In addition, although the overall workflow is more time-consuming than some simpler baselines, the measured detection time after training remains below one second per lumisection, which is in principle enough for online application. This timing should be regarded as a rough upper bound rather than as an optimized benchmark, since speed was not a main optimization target in this study and could likely be improved if needed.

At the same time, the final performance depends not only on the VAE itself but also on the entire post-processing chain. The workflow contains specific pattern rules, so the results depend not only on model quality but also on how the anomaly criteria are defined. For example, changing the required cluster size from 2×16 to 2×12 , or adopting a different anomaly pattern altogether, can lead to different flagged results. Several task-dependent thresholds also limit portability and generalization: although the present configuration is

effective for the current dataset and anomaly definition, the tuned settings may no longer remain appropriate for future datasets or different anomaly patterns. These limitations motivate the search for a more end-to-end alternative, which leads to the CLIP-based approach discussed in the following chapter.

Chapter 5

CLIP-Based Anomaly Detection

To reduce the dependence of the VAE-based workflow on manually chosen thresholds, a second methodological direction is explored in this chapter based on Contrastive Language-Image Pre-training (CLIP) [11]. Unlike the VAE pipeline, which converts reconstruction errors into anomaly decisions through several stages, a CLIP-based approach offers a more direct mapping from image input to class prediction. This is attractive for the present task because it does not require complex threshold tuning and can effectively leverage its transfer learning capability from large-scale pretraining, even when anomalous data are scarce. All the available real anomalous cases used here are taken from the VAE-based workflow described in Chapter 4.

5.1 CLIP and LoRA

A CLIP model consists of an image encoder and a text encoder trained jointly on large collections of image-text pairs. During pre-training, both images and their paired texts are mapped into a shared embedding space and optimized with a contrastive objective: matched image-text pairs are pulled closer, while mismatched pairs are pushed apart. After pre-training, the text encoder can embed class names or prompt templates, and predictions can be made by comparing these text features with the image features. In this way, CLIP turns image classification into a similarity-matching problem between visual and textual representations.

Low-Rank Adaptation (LoRA) provides a parameter-efficient way to adapt such a pre-trained model to a new task [12]. Instead of updating all weights of the backbone, LoRA freezes the original parameters and learns only low-rank update matrices for selected layers. In the few-shot setting, this reduces the number of trainable parameters, lowers optimization cost, and makes adaptation feasible even when the target dataset is small. Another practical advantage is that the learned low-rank updates can be merged into the original weights before inference, so the adapted model does not require an additional inference-time branch.

In standard CLIP usage, the text encoder is an essential part of the prediction interface, because the model compares image features with text prompts such as class names or short descriptions. This is not well matched to the anomaly types considered in this thesis. The anomalies studied here are defined by detector-specific spatial patterns in occupancy maps, such as extended single-layer deficits or cross-layer overlaps, and these patterns cannot be described naturally or precisely by a small set of prompt templates. For this reason, the present study does not use the text encoder as the main decision interface. Instead, it takes the pre-trained CLIP image encoder as the representation backbone and adapts it with LoRA on CMS occupancy images. The goal is to exploit the transferable visual features learned during large-scale pre-training while avoiding an artificial text formulation of detector anomalies.

5.2 Few-shot Learning

Because CLIP is explicitly designed to transfer pre-trained visual representations to downstream tasks with limited labeled data, the first experiment in this chapter follows a few-shot setting.

5.2.1 Pipeline

The few-shot pipeline is simple. A CLIP vision backbone is combined with a linear classification head, and LoRA is applied to the attention projection layers of the vision

transformer. The input preparation is the main task-specific part of the pipeline.

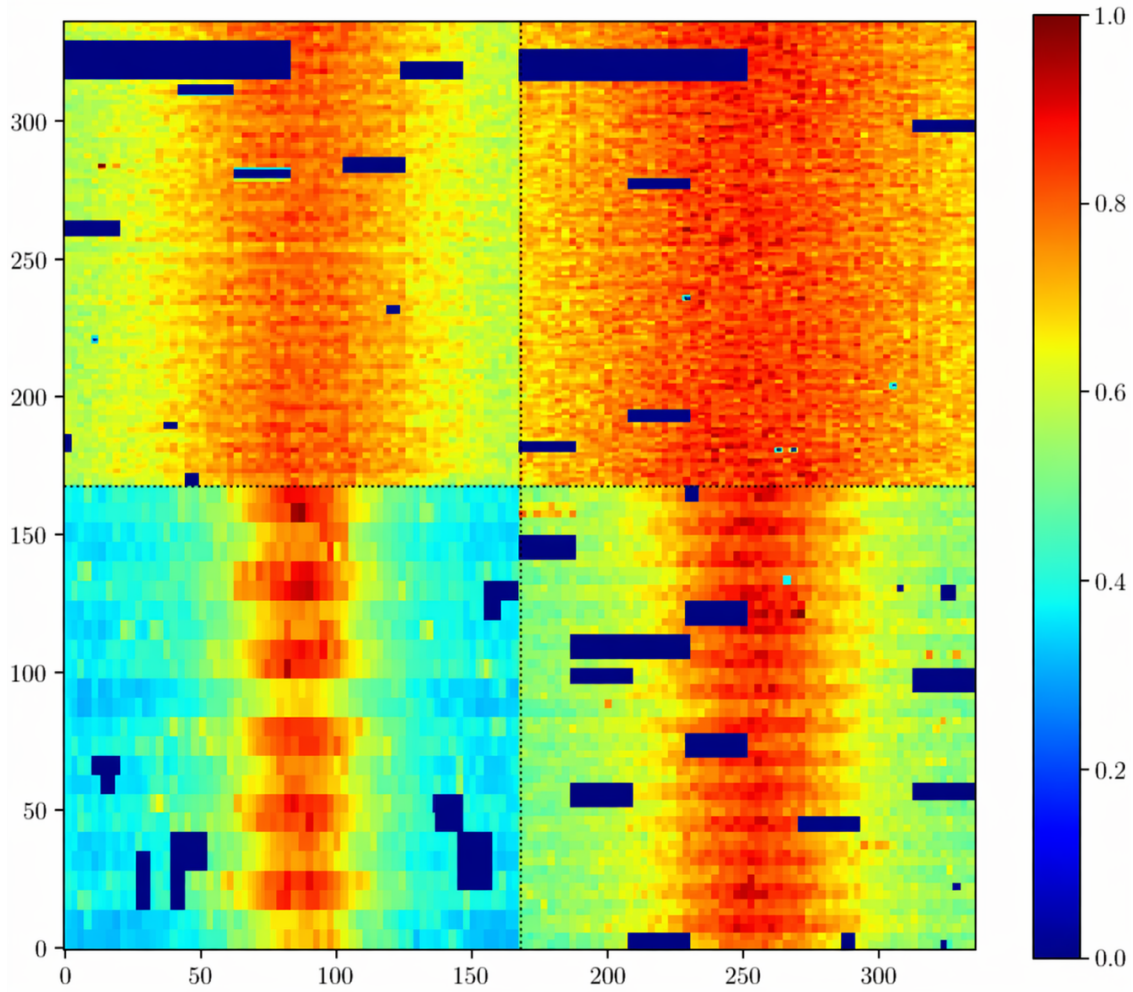


Figure 5.1: Preprocessing for CLIP few-shot learning.

As shown in Figure 5.1, the four BPix layers are first rebinned to the same size and arranged in a 2×2 layout to form one square composite image. Because CLIP expects a three-channel image rather than four detector maps, the composite is converted to RGB with a colormap after normalization to the $[0, 1]$ range. The resulting image is then processed by the CLIP image processor and passed to the CLIP vision encoder. In this few-shot study, training uses only 48 anomalous and 48 normal lumisections from Era C.

5.2.2 Results and Analysis

The accuracy is about 52%. This shows that the direct CLIP-based classifier has almost no effective discriminative power for the present task. It is only slightly above random guessing and is far below what would be needed for practical anomaly detection.

This outcome is not surprising. First, the 2×2 composite representation weakens the geometric relation between layers. In the VAE-based workflow, overlap across layers is introduced explicitly through rebinning and combination, whereas in the CLIP input the four layers are separated into different quadrants of one RGB image. As a result, the cross-layer information relevant to anomaly definition is no longer represented in a direct way. Second, the task itself is closer to fine-grained structural discrimination than to ordinary semantic recognition. The differences between normal and anomalous samples are often subtle, and such anomaly patterns are unlikely to be represented in CLIP’s natural-image pre-training corpus. As a result, the transferred representation does not provide a natural basis for detecting anomalous occupancy maps.

5.3 Training with More Data via Anomaly Injection

The weak performance in the few-shot setting suggests that the gap between CLIP pre-training and CMS occupancy maps cannot be overcome with only a very small number of labeled samples. To reduce this domain mismatch, a larger training set is constructed. Because the number of real anomalous lumisections is limited, anomaly injection is introduced to generate additional anomalous training examples.

5.3.1 Anomaly Injection

The anomaly-injection strategy is designed to generate synthetic anomalies that remain consistent with the anomaly definition introduced in Chapter 3. For a selected pair of layers, the layer with fewer rows is first taken as the base layer, and a valid injection region is

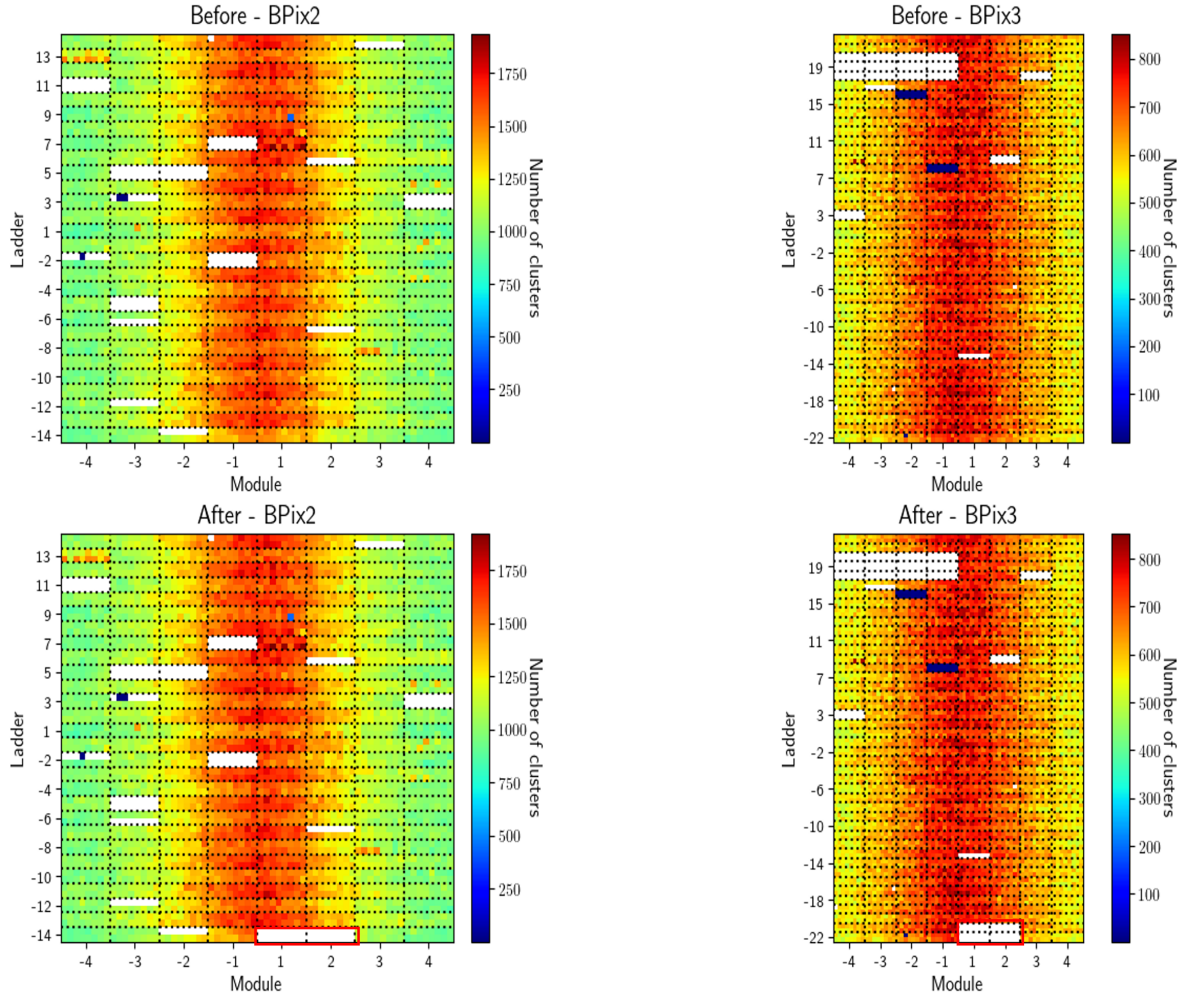


Figure 5.2: Example of anomaly injection on Era C.

chosen on the base map. The placement is aligned to the module grid, with a minimum size of 2×16 , and regions dominated by long-term zero areas are avoided. The selected anomalous region is then projected to the other layer so that the injected structures remain geometrically corresponding across layers.

After that, the selected module region is modified by a multiplicative factor corresponding to different anomaly types, such as complete suppression, reduced occupancy, or enhanced occupancy. After the anomalous block is inserted, multiplicative Gaussian noise is applied to all layers of the same lumisection, and a per-layer 99th-percentile clipping step is used during preprocessing to reduce the effect of bright outliers.

In this way, the injected samples preserve the overall occupancy pattern of real data while introducing controlled anomalous regions that are large enough and sufficiently structured to be relevant for training.

5.3.2 Results and Analysis

The same CLIP-based pipeline as in Section 5.2 is used here, but the training set is enlarged to 800 normal and 800 injected anomalous lumisections. Compared with the few-shot setting, the classification result improves substantially. On the real test set, the model reaches an accuracy of about 74.3%, which shows that increasing the amount of training data does help the CLIP vision encoder adapt to the occupancy-map domain.

However, the result is still below expectation. The classifier is clearly more capable of separating the two classes than in the few-shot experiment, but the number of false positives remains non-negligible and the overall result still falls short of a reliable detector-monitoring tool. This outcome is understandable. A transformer-based image encoder can benefit from more training data, so improvement over the few-shot setting is natural. At the same time, the two main difficulties identified in Section 5.2 still remain and cannot be easily overcome simply by increasing the size of the training set.

In principle, further enlarging the synthetic training set might improve the result again.

However, beyond a certain point this would also weaken the original motivation for using CLIP as a transferable pre-trained model. If large amounts of synthetic data are required for training, one could instead train a model designed directly for four-channel detector images from scratch, rather than adapting CLIP, although this would require a much larger training effort.

Chapter 6

Conclusion

6.1 Summary of Results

This thesis investigated CMS pixel-detector anomaly detection through two methodological directions. The first direction was a VAE-based workflow built around reconstruction error. This method produced results that remained broadly consistent with the previous NMF-based baseline while also identifying additional anomalous lumisections.

The second direction explored whether a CLIP-based model could reduce the dependence on hand-designed post-processing rules by learning a more direct classifier from images. In the few-shot setting, the result was close to random and showed that direct transfer from natural-image pre-training is not sufficient for this task. Enlarging the training set through anomaly injection improved the result substantially, indicating that the CLIP vision encoder can adapt to occupancy-map data when more task-relevant samples are provided. However, the performance still remained below the level required for a reliable monitoring tool. The main reasons are that the 2×2 image construction weakens the explicit representation of cross-layer overlap, and that the domain gap between CLIP pre-training and CMS occupancy maps is not easily removed.

Taken together, these results suggest that the VAE-based method is the stronger practical solution in this thesis, while the CLIP-based study serves as a useful exploration of the limits of transferable visual backbones in this setting.

6.2 Future Work

Several directions may improve upon the results presented here. First, the trained VAE could be used not only for anomaly detection but also for data generation. Since VAE can, in principle, sample new lumisections, it may provide a way to construct larger amounts of high-quality training data and thus support larger-scale supervised learning. Second, it would be natural to develop models that accept the four detector layers jointly in their original structured form. Such models could preserve cross-layer geometric relations more faithfully and may be better matched to the anomaly definitions used in this work. Third, instead of making only a hard anomalous-or-normal decision, future methods could output an anomaly score. This would return part of the final decision to human experts, reduce the risk of missed anomalies, and still lower the manual workload by allowing attention to be focused on the higher-scoring cases.

Bibliography

- [1] Lyndon Evans and Philip Bryant. Lhc machine. *Journal of Instrumentation*, 3(08):S08001, aug 2008.
- [2] The CMS Collaboration. The cms experiment at the cern lhc. *Journal of Instrumentation*, 3(08):S08004, aug 2008.
- [3] Virginia Azzolini, Broen van Besien, Dmitrijus Bugelskis, Tomas Hreus, Kaori Maeshima, Javier Fernandez Menendez, Antanas Norkus, James Fraser Patrick, Marco Rovere, and Marcel Andre Schneider. The Data Quality Monitoring software for the CMS experiment at the LHC: past, present and future. *EPJ Web Conf.*, 214:02003, 2019.
- [4] Vichayanun Wachirapusitan. Machine learning applications for data quality monitoring and data certification within cms. *Journal of Physics: Conference Series*, 2438:012098, 02 2023.
- [5] Anomaly detection for the CMS Pixel Barrel Data Quality with Machine Learning models. 2025.
- [6] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [7] Abhirami Harilal, Kyungmin Park, and Manfred Paulini. Anomaly detection based on machine learning for the cms electromagnetic calorimeter online data quality monitoring, 2024.

- [8] The Tracker Group of the CMS Collaboration. The cms phase-1 pixel detector upgrade, 2020.
- [9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.
- [10] Anomaly detection in data quality with machine learning models for the CMS pixel detector. 2026.
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [12] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models, 2024.