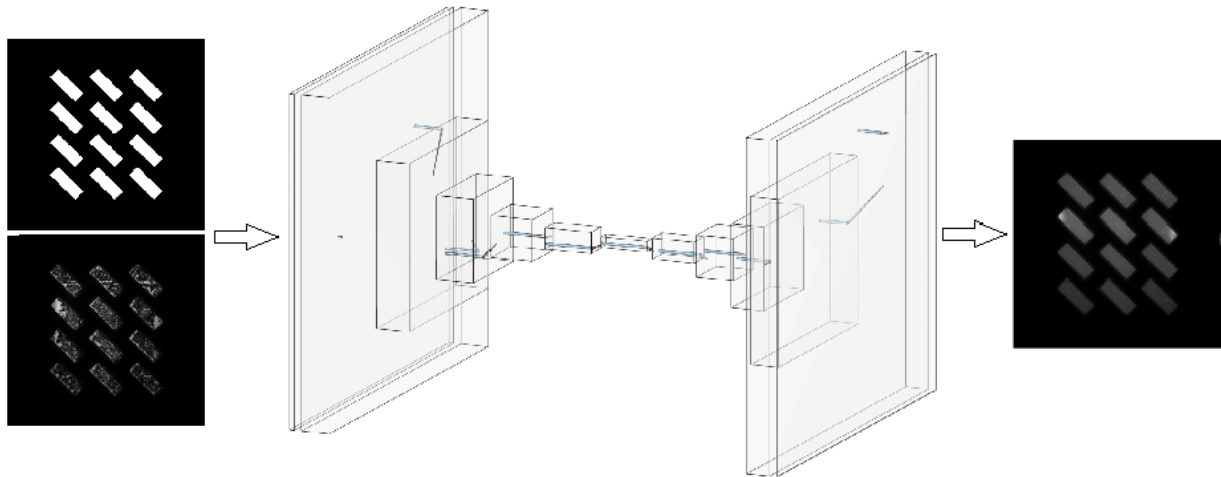




Unsupervised Machine Learning for Anomaly Detection in Electron Beam Powder Bed Fusion

Automating Quality Control of Electron Beam Melting layers

Master's thesis in Complex Adaptive Systems

KEIVAN HABIBI

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

MASTER'S THESIS 2026

Unsupervised Machine Learning for Anomaly Detection in Electron Beam Powder Bed Fusion

Automating Quality Control of Electron Beam Melting Layers

KEIVAN HABIBI



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Division of Subatomic, High-Energy and Plasma Physics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Unsupervised Machine Learning for Anomaly Detection in Electron Beam Powder
Bed Fusion
Automating Quality Control of Electron Beam Melting Layers
KEIVAN HABIBI

© KEIVAN HABIBI, 2026.

Supervisor: Martin Gardfjell, Colibrium Additive
Examiner: Ulf Gran, Department of Physics

Master's Thesis 2026
Department of Physics
Division of Subatomic, High-Energy and Plasma Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in L^AT_EX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Unsupervised Machine Learning for Anomaly Detection in Electron Beam Powder Bed Fusion

Automating Quality Control of Electron Beam Melting Layers

KEIVAN HABIBI

Department of Physics

Chalmers University of Technology

Abstract

Electron Beam Melting is used in high-stakes industries, such as in the aerospace and medical sectors, thus requiring rigorous quality control processes. As some builds consist of over 7,000 images, manual analysis is not feasible. Current approaches to counter these issues rely on heuristic-based methods that require large annotated datasets. This study suggests an unsupervised machine learning approach for anomaly detection, trained only on nominal data, removing the need for labelled defect examples. This work conducts an ablation study exploring different input modalities across four neural network models: an autoencoder and three image-to-image models. Results show that the autoencoder trained on reconstructing Layer images, and the image-to-image model taking Rake images together with CAD Masks perform best by a notable margin, where the image-to-image model achieves an F1 score of 98% on a baseline build containing both nominal and anomalous layers. The anomaly score behaviour remains consistent across multiple build cycles, suggesting the approach is scalable beyond a single build. However, validation across additional geometries and process parameters remains an important direction for future work.

Keywords: Electron Beam Melting, Anomaly Detection, Unsupervised Deep Learning, Image-to-Image Translation, Autoencoder, Additive Manufacturing, Ablation Study, Input Modality, In-Situ Monitoring, Thermal Imaging

Acknowledgements

I am extremely grateful to my supervisor Martin Gardfjell for continuous support and feedback throughout this thesis. Many thanks to Anders Snis for providing computational resources that significantly facilitated this work. Lastly, I would like to thank my parents for inspiring and supporting me through my education.

Keivan Habibi, Gothenburg, May 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AE	Autoencoder
ANN	Artificial Neural Networks
AM	Additive Manufacturing
CNN	Convolutional Neural Network
EBM	Electron Beam Melting
EB-PBF	Electron Beam Powder Bed Fusion
ELO	Electron-Optical
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Networks
IoU	Intersection over Union
I2I	Image-to-Image
LPT	Low-Pressure Turbine
ML	Machine Learning
M2L	Mask-to-Layer
PBF	Powder Bed Fusion
TN	True Negative
TP	True Positive
RM2L	Rake-and-Mask-to-Layer
R2L	Rake-to-Layer

Contents

List of Acronyms	ix
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Background	1
1.2 Problem Description	2
1.3 Aim and Purpose	2
1.4 Delimitations	3
1.5 Related Work	4
2 Theory	5
2.1 Additive Manufacturing	5
2.1.1 Powder Bed Fusion	5
2.1.2 Electron Beam Powder Bed Fusion	5
2.2 Geometric Camera Modeling and Calibration	6
2.3 Convolutional Neural Networks (CNN)	7
2.4 Encoder-Decoder Architectures	8
2.5 Image-To-Image Translation	8
2.6 Autoencoders	10
2.7 Loss function	10
2.7.1 Mean Absolute Error (L_1 Loss)	11
2.8 Evaluation Metrics for Imbalanced Datasets	11
3 Methods	13
3.1 Data acquisition and Pre-processing	13
3.2 Image Rectification and Alignment	14
3.3 Data Integration and Tiling	16
3.4 Network Architectures	17
3.4.1 Image-to-Image Translation Models	17
3.4.2 Undercomplete Autoencoder	18
3.5 Training Strategy and Implementation Details	19
3.6 Anomaly Scoring and Evaluation Framework	20
4 Results	23

4.1	Model Performance Comparison	23
4.2	Anomaly Detection on Build C	24
4.3	Generalization Across Build Cycles	24
4.4	Training and Validation Behaviour	27
5	Discussion	31
5.1	Combined Input and Autoencoder Performance	31
5.2	Ablation Study: The Role of Input Data	31
5.3	Geometric Data Imbalance and Anomaly Score Spikes	32
5.4	Overfitting and Architectural Considerations	33
5.5	Conclusion	33
	Bibliography	35

List of Figures

1.1	A titanium aerospace turbine blade manufactured via Electron Beam Melting.	1
1.2	LayerQam system mounted above the heat shield, side view.	2
1.3	The GE9X turbofan engine, developed by GE Aerospace for the Boeing 777X.	3
2.1	Electron Beam Powder Bed Fusion equipment. (a) Internal schematic showing the upper electron beam column housing the filament and electromagnetic control lenses, and the lower vacuum chamber containing the powder hoppers, rake mechanism, and incrementally descending build platform. (b) External view of the Spectra L machine.	6
2.2	Illustration of a convolution operation. A 3×3 kernel slides over a 7×7 input image, computing element-wise products at each position to produce a 5×5 feature map. The highlighted region (orange) shows the current kernel position, with the corresponding output value highlighted in blue.	8
2.3	The U-Net encoder-decoder architecture with skip connections. Reproduced from [1].	9
2.4	Encoder-decoder architecture illustrating the characteristic hourglass structure. The encoder (left) progressively reduces spatial resolution while increasing feature depth, compressing the input into a compact latent representation at the bottleneck. The decoder (right) symmetrically reconstructs the spatial resolution through upsampling. This architecture is shared across all models evaluated in this work.	9
3.1	Nominal layer image of a single layer showing six turbine blades being simultaneously melted within the electron beam melting (EBM) build chamber. The darker areas correspond to cold powder, while the brighter regions indicate higher temperatures. The melted geometry is surrounded by sintered powder.	14
3.2	Comparison of a nominal (top) and a non-nominal (bottom) layer from non-classified builds of a different geometry. The contrast between the two illustrates the type of deviations the models in this thesis are designed to detect.	15
3.3	Calibration plate featuring a symmetric grid of circles, used for Zhang’s camera calibration method [2].	16

4.1	Visual comparison of Precision, Recall, and F1 scores across model architectures (see Table 4.1 for exact values).	24
4.2	Integrated analysis of anomaly detection performance (top), physical geometry (middle), and volumetric cross-sectional area (bottom) throughout the total build height.	25
4.3	Classification results for Build C using the best performing model, RM2L.	26
4.4	Comparison of anomaly scores for the RM2L (Double Translation) model across several different build cycles. The dataset includes both nominal builds and those containing various types of industrial anomalies, illustrating how the model's reconstruction error fluctuates relative to the decision threshold depending on the specific build conditions and presence of defects.	27
4.5	Training and Validation Loss & IoU for all four models (1/2).	29
4.5	Training and Validation Loss & IoU for all four models (cont.).	30

List of Tables

3.1	Overview of the dataset and build characteristics.	13
3.2	Summary of the Machine Learning models implemented and evaluated in the study.	17
4.1	Performance metrics for anomaly detection across the different model architectures. The RM2L (Double Translation) architecture yields the highest F1 score by fusing CAD and Rake data evaluated on the calibration build (Build C). Note that threshold was selected to maximize F1 on this same dataset.	23

1

Introduction

This chapter introduces the industrial context and motivation behind the thesis, beginning with a background on Electron Beam Melting and its in-situ monitoring capabilities. The problem with current analysis methods is then outlined, followed by the aim and purpose of the work, its delimitations, and a review of related research in the field.

1.1 Background

Electron Beam Melting (EBM) is an Additive Manufacturing (AM) process that utilizes an electron beam to selectively melt metal powder in a layer-by-layer process [3]. Colibrium Additive, (formerly Arcam AB), pioneered the EBM technology, which is now employed globally across various high-stake industries. In the medical sector, EBM is used to manufacture orthopedic titanium implants, while the aerospace industry uses the technology for high-temperature applications such as turbine blades (see Figure 1.1). These applications place exceptionally high demands on quality assurance, increasing the importance of in-situ process monitoring.



Figure 1.1: A titanium aerospace turbine blade manufactured via Electron Beam Melting.

EBM belongs to the Powder Bed Fusion (PBF) family of AM processes, which operate through a cyclic, layer-by-layer procedure. Each cycle consists of three steps: (1) lowering the build platform by one layer thickness, (2) recoating the surface with fresh metal powder using a mechanical rake, and (3) selectively melting the powder using the electron beam. The geometry of each melted layer is determined by a CAD model that has been sliced into 2D cross-sections by in-house slicing software, these cross-sections are hereafter referred to as *CAD masks*. To monitor this process, cameras are typically mounted inside the build chamber to capture images of the powder bed surface. Colibrium Additive's proprietary monitoring solution, LayerQamTM (see Figure 1.2), captures infrared images of the top surface at two points per layer: one image after recoating (hereafter *Rake image*) and one after melting (hereafter *Layer image*). For a typical build of approximately 350 mm

height with a 0.09 mm layer thickness, this results in more than 7,000 images per build, meaning a substantial volume of data is generated that is in need of automated analysis. Machine learning-based approaches offer a promising alternative, enabling anomaly detection at scale without dependence on large, manually labeled datasets.

1.2 Problem Description

While LayerQam provides the necessary data for real-time monitoring, the analysis of these images remains a manual, time-consuming effort. Current attempts to automate this process have relied on heuristic-based image analysis using libraries such as OpenCV. While effective for certain tasks, these methods face three primary challenges: they require extensive annotated datasets for validation, they struggle to detect uncommon or novel defects where data is scarce, and they are difficult to scale due to the confidential nature of customer data. These limitations collectively motivate the need for an unsupervised approach that can operate without large, manually annotated datasets.



Figure 1.2: LayerQam system mounted above the heat shield, side view.

1.3 Aim and Purpose

The aim of this Master’s thesis is to investigate the potential of automating the analysis of in-situ layer-wise imagery in EBM using Machine Learning (ML). Specifically, the project explores the feasibility of implementing ML architectures that can identify process anomalies without relying on manual data annotation.

To address the challenges of data confidentiality and the high cost of labeling, the study focuses on a data-driven approach where labelling is performed through automated heuristics rather than manual annotation. A central part of this research is a systematic ablation study designed to quantify how different sources of contextual information influence the model’s predictive accuracy.

By establishing a correlation between input process parameters and the resulting thermal images, the project aims to:

- Develop and implement a reconstruction-based deep learning architecture using an Autoencoder framework for unsupervised anomaly detection.
- Develop and implement a context-aware Image-to-Image framework to predict nominal melt pool behavior from input process and design conditions.
- Conduct an ablation study to evaluate the individual and combined contributions of geometric design data (CAD) and physical process state (Rake) as inputs, comparing three configurations: CAD only, Rake only, and CAD + Rake combined.
- Evaluate the ability of these models to detect deviations by analyzing the discrepancy (residual error) between the model's predictions and the actual process data.
- Assess the viability of these automated approaches as a robust and scalable alternative to current manual inspection routines.

The project will conclude by discussing the reliability and robustness of the ML-based framework compared to traditional heuristic-based methods, specifically focusing on its ability to generalize across different geometries and its independence from manual labeling.

1.4 Delimitations

The study is delimited to process data from a specific build geometry, low-pressure turbine blades (LPT-blades) in titanium aluminide for the GE9X aircraft engine as seen in Figure 1.3. Although the underlying ML principles are general, the performance of the model on significantly different geometries or materials is outside the scope of this thesis.



Figure 1.3: The GE9X turbofan engine, developed by GE Aerospace for the Boeing 777X.

This research focuses on global anomaly detection, meaning identifying whether a layer or region is deviating from the norm, rather than high-precision pixel-wise segmentation. This choice is motivated by the stochastic nature and low signal-to-noise ratio of EBM melt images. Therefore, the global approach is adopted, as this

allows the network to prioritize the learning of physical trends rather than localized noise.

1.5 Related Work

Prior research into anomaly detection for Electron Beam Melting (EBM) has explored various data modalities to identify print defects. Several studies have focused on machine-produced time-series data [4] or the analysis of electron emission levels [5]. Both of these approaches utilize deep encoder-decoder architectures to leverage reconstruction error as a metric for anomaly quantification. Additionally, transfer learning has been applied to electron-optical (ELO) images to classify anomalies into predefined categories [6].

While the application of Machine Learning within the EBM domain is an evolving field with significant potential [7], a distinct research gap remains regarding the use of high-resolution, in-situ optical monitoring. To the best of the author’s knowledge at the time of writing, there is a lack of literature exploring deep generative or reconstruction-based models specifically applied to layer-wise powder bed imagery in EBM.

The utility of using residual errors from generative networks has been well-documented in computer vision [8, 9]. However, there is a lack of systematic research quantifying the added value of combining digital design data (CAD) with physical process states (Rake) within an image-to-image framework for EBM. This highlights the need for a comparative evaluation to determine the optimal data configuration for robust in-situ monitoring.

2

Theory

This chapter presents the theoretical foundations underpinning the methodology developed in this thesis. It covers the relevant aspects of additive manufacturing and electron beam melting, followed by the image processing and deep learning concepts central to the proposed anomaly detection approach, including encoder-decoder architectures, image-to-image translation, and autoencoders. The chapter concludes with the loss functions and evaluation metrics used throughout the study.

2.1 Additive Manufacturing

Additive Manufacturing (AM) is defined as the process of joining materials to make parts from 3D model data, usually layer upon layer [10], in contrast to traditional subtractive manufacturing where material is progressively removed. The ISO/ASTM 52900 standard categorises commercially available AM systems into seven distinct process categories [10, 11], of which Powder Bed Fusion (PBF) is relevant to this work.

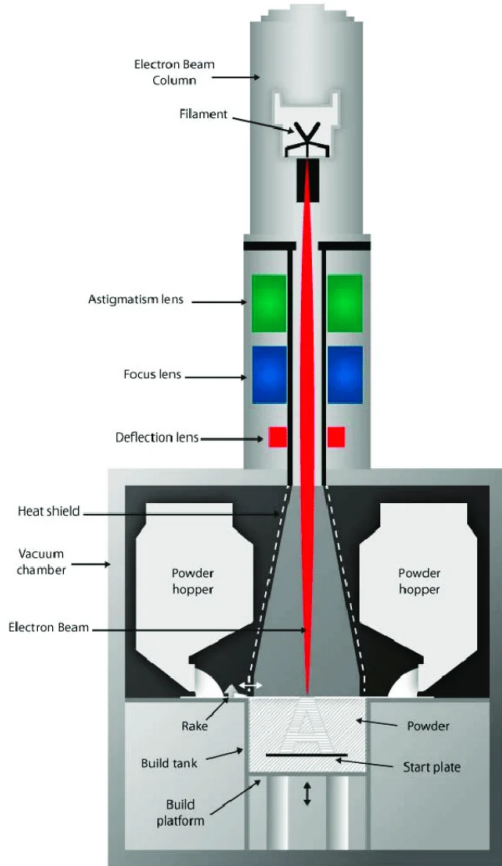
2.1.1 Powder Bed Fusion

PBF is the process of continuously feeding powder to a platform where each cycle of powder is used to melt (or fuse) the surface layer in an enclosed chamber [3, 10]. This loop is continued until the entire build is finished. More specifically, it is a process using one of three heat sources: a laser beam, an electron beam, or a plasma arc beam where respective methods have their strengths in different areas. The thermal energy from said heat source selectively melts each layer when it comes into contact with the powder, and once the layer has cooled and thus fused, a new layer of powder is provided as the building plate is lowered iteratively [3].

2.1.2 Electron Beam Powder Bed Fusion

Electron Beam Powder Bed Fusion (EB-PBF), or (EBM), combines Powder Bed Fusion together with an Electron Beam. EBM is an approach where electrons are controlled in a vacuum together with electric and magnetic fields to create a beam that can be controlled. The high-velocity of the electron beam interacts with metals, which when colliding in a focused manner produces enough heat to selectively fuse metal powder [3]

As illustrated in Figure 2.1, the rake component evenly distributes the powder for each layer, thereafter the electron beam with the help of the three lenses is positioned and controlled and the top layer is consequently melted inside the vacuum chamber. This is done iteratively, meaning the build platform is successively lowered after each layer has successfully been melted by the electron beam, and ultimately the rake distributes a fresh layer of powder onto the platform for the next cycle.



(a) Schematic representation of an EB-PBF machine.



(b) The Spectra H EB-PBF machine used in this work.

Figure 2.1: Electron Beam Powder Bed Fusion equipment. (a) Internal schematic showing the upper electron beam column housing the filament and electromagnetic control lenses, and the lower vacuum chamber containing the powder hoppers, rake mechanism, and incrementally descending build platform. (b) External view of the Spectra L machine.

2.2 Geometric Camera Modeling and Calibration

The standard model for mapping three-dimensional coordinates to a two-dimensional image plane is the pinhole camera model [12]. The projection of a 3D point $\mathbf{M} = [\mathbf{X}, \mathbf{Y}, \mathbf{Z}, 1]^T$ to a 2D pixel coordinate $\mathbf{m} = [\mathbf{u}, \mathbf{v}, 1]^T$ is mathematically defined as:

$$s\mathbf{m} = A[R|t]\mathbf{M} \quad (2.1)$$

where s is an arbitrary scale factor, A represents the intrinsic matrix (containing focal length and principal point), and $[R|t]$ denotes the extrinsic parameters (rotation and translation) which define the camera's spatial orientation and position.

In industrial environments such as EBM, estimating these parameters is often challenging due to limited space and non-perpendicular camera mounting. To address this, Zhang's calibration method [2] is implemented. The technique utilizes a planar target with known geometry to solve the projection equations. As stated in Zhang's publication [2], a key simplification in this method is the assumption that the calibration plane lies at $Z = 0$ in the coordinate system. Using this, the relationship simplifies to a planar homography (H):

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = A \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \quad (2.2)$$

This homography matrix H is a 3 x 3 matrix that directly maps points from the physical build plate to pixels in the image. Furthermore, to account for geometric asymmetries, the standard pinhole model is extended using a tilted sensor model [13]. This introduces additional parameters, τ_x and τ_y , which describe the rotation of the image plane. These parameters are particularly relevant in industrial setups where a camera observes a scene through a protective viewport at an oblique angle. In such cases, the model accounts for the non-uniform refraction introduced by the glass, while also allowing mathematical flexibility for compensation of minor mechanical deviations or thermal-induced shifts in the system geometry.

2.3 Convolutional Neural Networks (CNN)

The analysis of complex images produced during the EBM process requires a thorough method for pattern recognition. Consequently, Convolutional Neural Networks (CNNs) are utilized. As described by O'Shea and Nash [14], CNNs are a class of biologically inspired artificial neural networks (ANNs) specifically architected to solve difficult image-driven recognition tasks.

What distinguishes standard ANNs from CNNs is that CNNs employ convolutional layers that use mathematical filters (kernels) to extract spatial hierarchies and features (see Figure 2.2), such as the edges of a melt pool or the texture of the powder bed. However, it is important to note that while CNNs are generally perceived as translation-invariant, recent research suggests otherwise [15]. It is not an architectural feature, but rather a feature that the network must learn through exposure to diverse data. This shows the importance of introducing a broad dataset for training, where the model learns to interpret the different shapes of features in different segments.

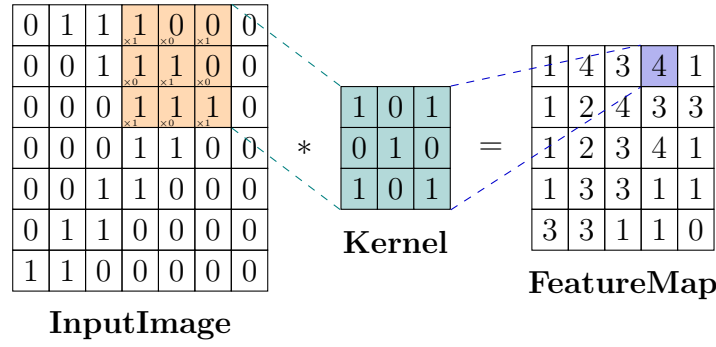


Figure 2.2: Illustration of a convolution operation. A 3×3 kernel slides over a 7×7 input image, computing element-wise products at each position to produce a 5×5 feature map. The highlighted region (orange) shows the current kernel position, with the corresponding output value highlighted in blue.

2.4 Encoder-Decoder Architectures

Building on the principles of Convolutional Neural Networks described in 2.3, the encoder-decoder architecture serves as a foundational framework for mapping high-dimensional input data to a specific output representation through feature abstraction and spatial reconstruction. Characterized by a symmetrical hourglass shape (see Figure 2.4), it consists of two primary components:

- **Encoder:** Performs hierarchical feature extraction using convolutional kernels to gradually reduce spatial resolution while increasing the number of feature channels, culminating in a low-dimensional bottleneck or latent space.
- **Decoder:** Projects these latent features back into a high-resolution spatial domain using upsampling or transposed convolutions.

While this approach was popularized by the U-Net framework [1] (illustrated in Figure 2.3), which utilizes skip-connections to preserve spatial details, certain implementations deliberately omit these connections. Removing these direct paths forces the network to process all information through the restricted latent bottleneck, acting as a regularization mechanism that prevents the simple duplication of low-level pixel noise in favour of learning generalized representations.

2.5 Image-To-Image Translation

Building on the encoder-decoder framework introduced in Section 2.4, Image-to-Image translation involves the use of neural networks to map data between two different domains. This field has been established through frameworks such as Pix2Pix [16] and SPADE [17], which typically utilize conditional generative models to produce outputs that are semantically consistent with the input data.

While Generative Adversarial Networks (GANs) are frequently employed to enhance

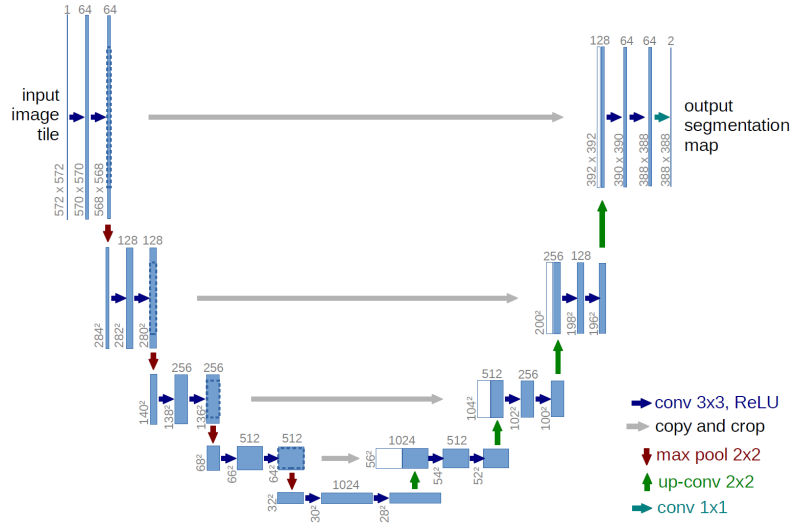


Figure 2.3: The U-Net encoder-decoder architecture with skip connections. Reproduced from [1].

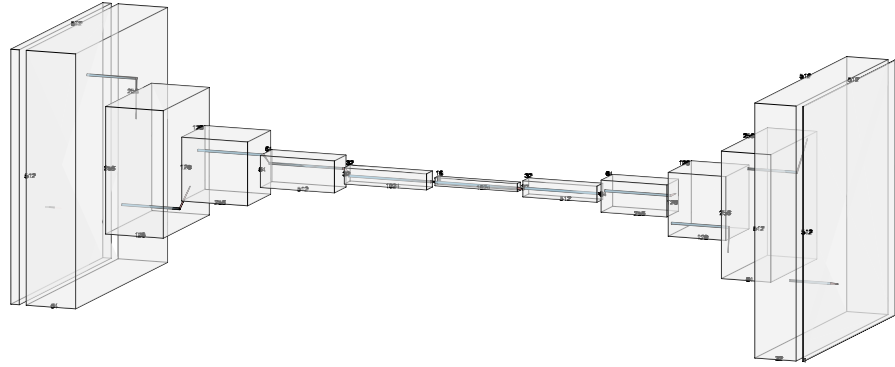


Figure 2.4: Encoder-decoder architecture illustrating the characteristic hourglass structure. The encoder (left) progressively reduces spatial resolution while increasing feature depth, compressing the input into a compact latent representation at the bottleneck. The decoder (right) symmetrically reconstructs the spatial resolution through upsampling. This architecture is shared across all models evaluated in this work.

visual realism, the translation task can also be formulated as a dense supervised regression problem. By utilizing pixel-wise loss functions, the model is optimized for numerical proximity to the ground truth, ensuring physical and numerical consistency between the input features and the generated output.

In this work, the translation is formulated as a regression problem rather than an adversarial one, optimising directly for pixel-wise proximity to the ground truth without a discriminator component.

In the context of anomaly detection, the core idea is that the model is trained exclusively on nominal process data. By doing so, the model learns to predict what a melt pool should look like under normal conditions, given the input context. When an anomalous layer is encountered during inference, the model still produces a prediction consistent with nominal behaviour since this is the only distribution it has learned. The resulting residual between the prediction and the actual captured image is therefore elevated, forming the basis for anomaly detection.

2.6 Autoencoders

An autoencoder is a neural network architecture designed for representation learning where the input and target output are identical. Formally, the model consists of an encoder function $h = f(x)$ and a decoder function $r = g(h)$, where r is the reconstruction of the input x .

As noted by Goodfellow et al. [18], the objective is typically to create an under-complete architecture where the dimension of the latent representation h is smaller than that of x . By forcing the data through a restricted bottleneck, the network is encouraged to prioritize the most representative features of the training distribution. When trained exclusively on good data, the model lacks the capacity to accurately reconstruct anomalous features. Consequently, these anomalies result in a significantly higher reconstruction error ($\|x - r\|^2$), allowing them to be identified through pixel-wise residual analysis.

2.7 Loss function

In the field of deep learning, the loss function, or criterion, acts as the objective function that the model aims to minimize during training. It quantifies the discrepancy between the network's output and the target ground truth. For reconstruction tasks such as image-to-image translation and autoencoders, the choice of loss function is critical, as it directly influences the sharpness and quality of the reconstructed output.

2.7.1 Mean Absolute Error (L_1 Loss)

The L_1 loss, or Mean Absolute Error, measures the average of the absolute difference between each pixel in the predicted image $\hat{y}$ and the ground truth image y . It is defined as:

$$L_1 = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.3)$$

Where:

- n is the total number of pixels in the image.
- y_i represents the value of the i -th pixel in the ground truth image.
- $\hat{y}_i$ represents the value of the i -th pixel in the predicted image.

A key characteristic of L_1 is that it treats all error linearly. In computer vision, this often results in sharper reconstructions compared to other metrics. Because L_1 does not disproportionately penalize large outlier, it is considered more robust to noise and is widely used in I2I tasks to preserve high-frequency details [16].

2.8 Evaluation Metrics for Imbalanced Datasets

In anomaly detection, the datasets are inherently imbalanced, as anomalous samples are typically heavily outnumbered by nominal ones. In such scenarios, relying on overall Accuracy can be highly misleading as a model who predicts every layer as good would achieve a near-perfect accuracy while failing its intended task of detecting anomalies [18].

Instead, the performance of the model is quantified using Precision, Recall, and the F1-score. These metrics consists of True Positives (TP), False Positives (FP), and False Negatives (FN), where a "Positive" represents a detected anomaly.

Precision measures the reliability of the alarms generated by the model, defined as the ratio of correctly identified anomalies to the total number of predicted anomalies:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the model's ability to discover all actual anomalies present in the dataset:

$$Recall = \frac{TP}{TP + FN}$$

There is typically a trade-off between these two metrics. Maximizing Recall often leads to an increase of false alarms (low precision), which would evidently result in alarm fatigue in our case. To find the optimal threshold, the F1-score is applied. The F1-score is the harmonic mean of Precision and Recall, providing a single metric that requires both to be reasonably high to achieve a good score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3

Methods

This chapter describes the full methodology developed to enable automated anomaly detection in electron beam melting. It covers the data acquisition pipeline and pre-processing steps required to produce spatially aligned, model-ready inputs, followed by a description of the four neural network architectures implemented in the ablation study. The training strategy, hyperparameter configuration, and hardware setup are then detailed, before concluding with the anomaly scoring and evaluation framework used to assess and compare model performance across multiple build cycles.

3.1 Data acquisition and Pre-processing

The LayerQamTM system within the EBM machine captures two types of images: a Rake image and a Layer image. The Rake image records the powder distribution across the build area prior to melting, while the Layer image captures the state of the powder bed after the electron beam has completed the melting process. These raw images are captured at a resolution of 4000×4000 pixels. To enable pixel-wise analysis and Image-To-Image translation, these data sources must be transformed into a unified spatial domain. In this thesis, data from different builds is used. The builds are presented in Table 3.1, with a corresponding description of the builds characteristics. Important to note is that only build A is utilized for the model training, whereas the rest are solely present in the evaluation phase.

Table 3.1: Overview of the dataset and build characteristics.

Build ID	Ref.	Description
0933	A	Nominal build used for training.
0932	B	Oversaturated full build.
0934	C	Nominal until 320 mm. Thereafter quality degrades significantly.
0935	D	Nominal until 50 mm. Thereafter metallization on glass causes images to go completely dark.
0953	E	Nominal build, but with oversaturation in geometrical transition areas at 50 mm and 300 mm.
1010	F	Nominal build throughout the full height.

Figure 3.1 shows a layer image of six turbine blades taken with the LayerQamTM system. The layer represents a cross-section of the full turbine blade geometry shown in Figure 1.1 and Figure 4.2b. The data that this thesis is based upon is confidential and is thus not included in this thesis. The images shown are therefore not used for the model training and evaluation, however they are considered highly representative.

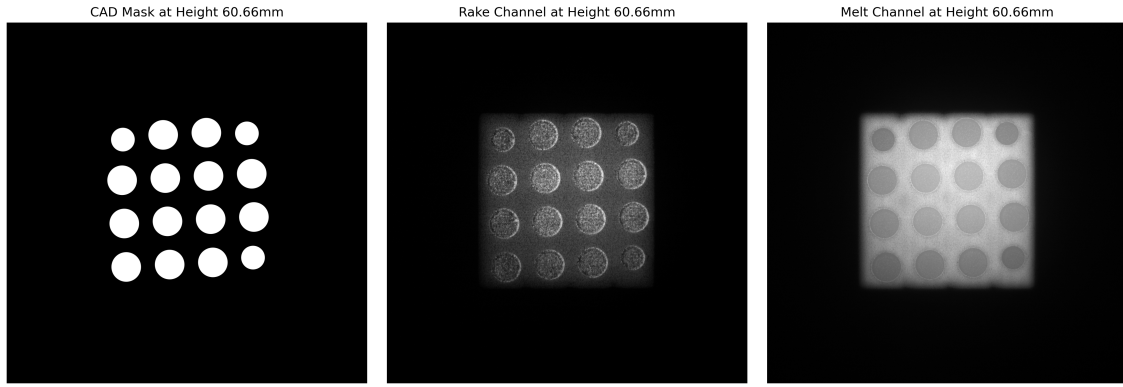


Figure 3.1: Nominal layer image of a single layer showing six turbine blades being simultaneously melted within the electron beam melting (EBM) build chamber. The darker areas correspond to cold powder, while the brighter regions indicate higher temperatures. The melted geometry is surrounded by sintered powder.

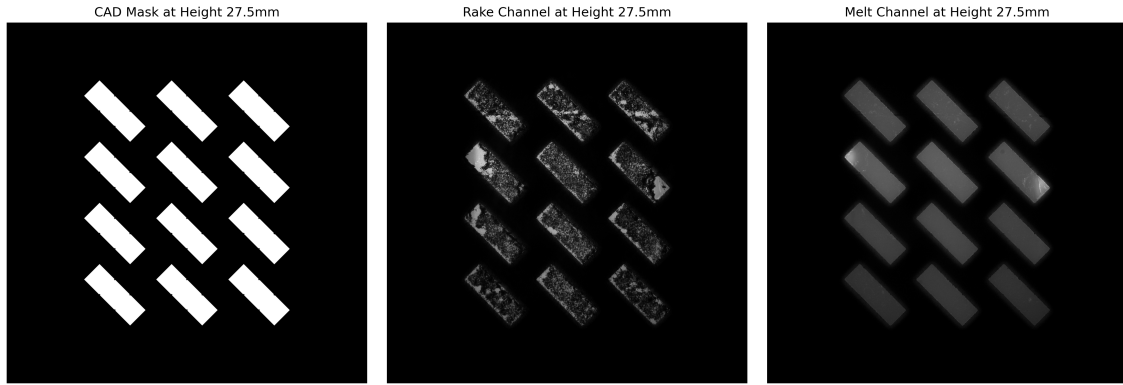
To further illustrate the difference between nominal and non-nominal process behaviour, Figure 3.2 presents a comparison of a nominal and a non-nominal layer, each displaying the CAD mask, Rake image, and Layer image side by side. Note that both examples originate from non-classified builds of a different geometry due to data confidentiality constraints, but are considered representative of nominal and non-nominal layer behaviour.

3.2 Image Rectification and Alignment

To align the different sensors, a physical calibration plate (see Figure 3.3) was placed on the EBM build plate. By detecting the control points on this pattern, the trans-



(a) Nominal layer. From left to right: CAD mask, Rake image, and Layer image. The powder distribution is uniform and the melt region closely follows the CAD mask geometry.



(b) Non-nominal layer at a height of 27.54 mm. From left to right: CAD mask, Rake image, and Layer image. Anomalies caused by uneven powder distribution due to rake deflection are visible in both the Rake and Layer channels.

Figure 3.2: Comparison of a nominal (top) and a non-nominal (bottom) layer from non-classified builds of a different geometry. The contrast between the two illustrates the type of deviations the models in this thesis are designed to detect.

formation between the physical world ($\mathbf{M}$) and the camera pixels ($\mathbf{m}$) was calculated.

Using the relationship defined in Equation 2.2, a homography matrix H was estimated. This matrix was used to perform a perspective warp on all raw images, effectively "un-projecting" the camera's oblique view into a top-down, rectified perspective. This step was critical to ensure that the Rake and Thermal domains were spatially synchronized, which is a prerequisite for the residual analysis used to identify process anomalies. Following rectification and cropping to the active build area, the resulting images have a resolution of 2500×2500 pixels.

The CAD masks are generated independently using in-house software and exists in a separate coordinate space from the LayerQamTM system. However, by utilizing the known physical dimensions of the calibration plate and the calculated spatial

resolution per pixel, these masks were transformed and scaled to align precisely with the rectified Rake and Thermal images.

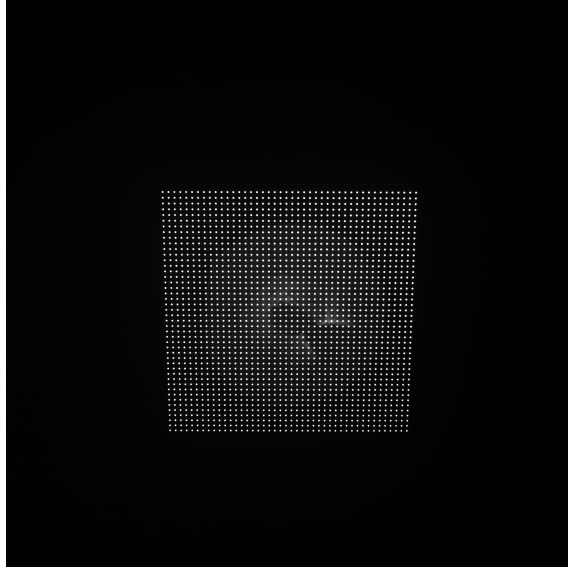


Figure 3.3: Calibration plate featuring a symmetric grid of circles, used for Zhang’s camera calibration method [2].

3.3 Data Integration and Tiling

For each build layer, the corresponding Thermal, Rake, and CAD images are concatenated into a single multi-channel data structure. Since the source images are single-channel grayscale, this integration allows for efficient data retrieval during model training without information loss. These combined datasets are stored as NumPy arrays for optimized I/O performance.

To address hardware memory constraints, specifically GPU VRAM limitations, a tiling approach was implemented. High-resolution images, such as the rectified 2500×2500 frames, require significant memory for both input storage and layer activations during the forward and backward passes of training. Attempts to process these images at full scale would render the use of mini-batch training computationally infeasible.

By subdividing the images into a grid of 512×512 tiles, the model can maintain high-resolution feature extraction while allowing for a sufficient batch size. This is critical for the stability of the gradient descent process and ensures that fine-grained thermal details are preserved, which would otherwise be lost if the entire build plate were downsampled to fit the model’s memory constraints.

3.4 Network Architectures

This section details the architectures of the proposed neural networks, outlining the specific implementation choices and the underlying rationale for their design. All models were implemented in Python using the PyTorch deep learning framework. To systematically evaluate the impact of geometric and process context on anomaly detection, a comparative **ablation study** was conducted using four distinct models, as summarized in Table 3.2.

The core of the study consists of three Image-to-Image (I2I) translation frameworks designed to predict the nominal melt pool layer. The primary model, **RM2L**, utilizes both the Rake and CAD-mask layers as dual-channel input. To isolate the contribution of each modality, two additional I2I models were developed: **R2L**, which relies solely on the Rake imagery, and **M2L**, which uses only the CAD-derived mask. Finally, a baseline Autoencoder (**AE**) is implemented to perform unsupervised reconstruction of nominal melt pool layers, serving as a benchmark for evaluating the predictive I2I translation approach.

Table 3.2: Summary of the Machine Learning models implemented and evaluated in the study.

Model Task	Architecture Type	Abbreviation
Rake-to-Layer	Image-to-Image translation	R2L
Mask-to-Layer	Image-to-Image translation	M2L
Rake-and-Mask-to-Layer	Image-to-Image translation	RM2L
Layer-to-Layer	Autoencoder (Baseline)	AE

3.4.1 Image-to-Image Translation Models

The Image-to-Image (I2I) models employ a deep encoder-decoder topology inspired by the U-Net architecture. However, to enforce a strict informational bottleneck, skip connections are deliberately omitted. In the context of anomaly detection, allowing identity mapping via skip connections would enable the network to directly copy input defects to the output. Instead, these models are forced to perform domain translation (and sensor fusion for RM2L) entirely through a compressed latent space.

The networks accept a 512×512 spatial input, with the channel depth dynamically adapted for each ablation model (two channels for RM2L, one for M2L and R2L). The data is processed through a deep, five-stage encoder utilizing custom *DoubleConv* blocks (consecutive 2D convolutions, Batch Normalization, and ReLU activations) and 2x2 Max Pooling for spatial reduction. As spatial dimensions are halved, channel depth progressively doubles from 64 to 1024.

At the deepest point of the network, the spatial resolution is drastically reduced to 16×16 pixels. This extreme compression acts as the primary chokepoint, forcing

the network to learn a highly efficient semantic representation of the physical geometry. To regularize this latent space and prevent overfitting, a Spatial Dropout layer (*Dropout2d* with $p = 0.3$) is applied to the bottleneck features.

Following the bottleneck, a symmetric five-stage decoder reconstructs the latent representation back to the original 512x512 resolution. Each decoding stage utilizes *ConvTranspose2d* for spatial upsampling followed by a *DoubleConv* block to refine the reconstructed feature maps. The final output block reduces the channel depth to 32 before applying a 1×1 convolution with a Sigmoid activation, yielding the single-channel predicted nominal melt pool. During inference, this prediction is compared against the actual captured melt pool image to compute the residual error, forming the basis of the anomaly alarm.

It is worth emphasising that all I2I models are trained exclusively on nominal data from Build A (3.1). The anomaly detection capability therefore does not rely on the model having seen defective examples, but rather on the assumption that a model trained on nominal behaviour will fail to accurately predict anomalous appearances, thus producing a high residual error because the anomaly lies outside the learned distribution.

3.4.2 Undercomplete Autoencoder

To establish a baseline and isolate the value of the contextual CAD and Rake data, a standard Undercomplete Autoencoder (AE) was implemented. Unlike the I2I models, the AE takes only a single-channel input: the actual captured melt pool image.

While the macro-architecture of the AE shares the same structural design as the I2I models, where it utilizes a deep five-stage encoder, a strict informational bottleneck with Spatial Dropout, and a symmetric decoder without skip connections, its internal capacity was deliberately scaled down. The I2I models require massive channel depth (peaking at 1024 channels) to successfully map complex geometric relationships across entirely different sensor domains.

Conversely, an Autoencoder operating within the same domain (Image $\rightarrow$ Image) is highly susceptible to memorization. If given excessive capacity, the network could theoretically act as an identity function, passing the input directly to the output without learning the underlying data distribution. To enforce a strict undercomplete constraint, the AE starts with just 16 channels and progressively increases to 256, before aggressively compressing the feature maps down to a mere 32 channels at the 16x16 bottleneck.

Like the I2I models, the AE is trained exclusively on nominal data from Build A. Because the network is heavily undercomplete, meaning the bottleneck has a drastically lower dimensionality than the input space, it cannot simply memorize the data. Instead, it is forced to learn a compact lower-dimensional representational of nominal melt pool characteristics. Anomalies, which deviate from this learned

nominal representation, fail to pass through the bottleneck accurately, resulting in a high reconstruction error (residual) that triggers the anomaly alarm.

3.5 Training Strategy and Implementation Details

To evaluate the proposed architectures and establish a robust ablation study, four distinct models were trained: the Undercomplete Autoencoder (AE), Mask-to-Layer (M2L), Rake-to-Layer (R2L), and the dual-input Double Translation model (RM2L). To ensure a rigorous and fair comparison, all models were trained using identical hyperparameters, with deviations made only to accommodate specific architectural constraints and memory footprints. The core training parameters shared across the models were:

- **Optimizer:** AdamW (Weight Decay: 1×10^{-4})
- **Initial Learning Rate:** 5×10^{-5}
- **Learning Rate Scheduler:** ReduceLROnPlateau (Factor: 0.5, Patience: 3, Min LR: 1×10^{-7})
- **Early Stopping:** Patience of 10 epochs (Minimum delta: 1×10^{-5})
- **Criterion:** L1 Loss (Mean Absolute Error), chosen to encourage sharper reconstructed boundaries and reduce sensitivity to outliers.
- **Epochs:** 50
- **Precision:** Automatic Mixed Precision (AMP) via `torch.amp` to optimize memory utilization and computational throughput.
- **Workers:** 4
- **Data Augmentation:**
 - Spatial Flips: Horizontal and Vertical (50% probability)
 - Gaussian Noise ($\sigma = 0.1$)
 - Brightness Variation Factor: 0.1
 - Coarse Dropout Size: 10% of spatial dimensions*

* *Coarse Dropout* refers to the application of black masking boxes on the input images. To prevent the models from learning corrupt geometric foundations, this noise was applied strictly to the physical sensor data (the rake images and the actual melt pool images) while leaving the binary CAD masks unaltered. During training, each applicable input tile has a 50% probability of undergoing spatial dropout. When triggered, the image is conceptually divided into a grid where the size of each dropout box is 10% of the image dimensions (e.g., 51×51 pixels for a 512×512 tile). Each of these grid patches then has an independent 30% probability of being blacked out. Statistically, this means half of the dataset encounters masking, and in those instances, roughly 30% of the image area is obscured by localized black patches.

Training progress was monitored using both the loss value and the Intersection over Union (IoU), defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (3.1)$$

where A is the set of predicted foreground pixels and B is the set of ground truth foreground pixels. IoU provides an interpretable measure of spatial overlap between

the predicted and target output, complementing the loss value during training monitoring. For IoU computation, the predicted output is binarized using a fixed threshold of 0.5 and compared against the corresponding CAD mask, which serves as the binary ground truth representing the expected melt area. Since the CAD mask is already binary, only the predicted output requires binarization.

Due to differing memory requirements, the batch sizes were adjusted between the architectures. The computationally lightweight AE utilized a batch size of 32, whereas the significantly deeper, multi-channel I2I models (M2L, R2L, RM2L) were restricted to a batch size of 16 to prevent memory overflow on the GPU.

Prior to training, the dataset was filtered using the aforementioned tiling algorithm (Section 3.3). To ensure the models learned meaningful features, all completely empty spatial tiles (where the corresponding CAD image was strictly black) were discarded. This filtration resulted in a refined dataset of 64,505 informative tiles. A standard 90/10 split was applied, allocating 58,055 tiles purely for training and 6,450 tiles for validation.

Since tiling is applied prior to the split and tiles are shuffled randomly, validation tiles may originate from layers that are also represented in the training set. This means the validation set measures reconstructions quality on unseen tile regions rather than entirely unseen layers. A strictly layer-wise split would provide a more conservative generalization estimate and is left as a direction for future work.

Hardware Specifications:

- **CPU:** Intel(R) Xeon(R) W-2235 CPU @ 3.80GHz
- **GPU:** NVIDIA RTX A4000 (16 GB VRAM)
- **RAM:** 64.0 GB

3.6 Anomaly Scoring and Evaluation Framework

Following the training phase, an initial calibration and evaluation phase is conducted to establish a classification threshold and identify the most capable architecture. For this purpose, **Build C (0934)** is utilized as the primary calibration baseline. As detailed in Table 3.1, this build provides a highly representative sequence of both nominal and anomalous thermal behaviours within the exact same geometric environment.

Inference is initially executed exclusively on Build C across all four trained models. During this process, the models output a predicted nominal representation of the melt pool for each processed tile. To determine the anomaly score for a full physical layer, the L1 residual error is calculated pixel-by-pixel. Because the models operate on overlapping tiles, these local residuals are spatially reconstructed into a global 2500×2500 anomaly map. In regions where tiles overlap, the maximum residual value is retained to ensure no defect signal is suppressed. The layer’s overall anomaly score is defined as the global maximum error within this reconstructed spatial map:

$$Score = \max_{i,j} (|X_{i,j} - \hat{X}_{i,j}|) \quad (3.2)$$

where i and j represent the spatial pixel coordinates of the full layer.

The maximum operator was selected over alternatives such as percentile-based aggregation because empirical results showed stable and consistent anomaly separation across build cycles. Exploring alternative aggregation strategies, such as high-percentile scoring, represents a natural direction for future optimization.

For each of the four models, an optimal decision threshold that distinguishes nominal layers from anomalous layers is obtained by maximizing the F1-score on the Build C dataset. Following this calibration phase, the best-performing model is identified.

The threshold is intentionally kept fixed across all subsequent builds to simulate a realistic deployment scenario, where a pre-calibrated model is applied to unseen manufacturing conditions without access to ground truth labels. To evaluate generalisation and understand how this architecture behaves under different manufacturing conditions, this single superior model along with its locked threshold is subsequently applied across all remaining unseen test builds (Builds B, D, E, and F). If the anomaly score for a layer exceeds the established threshold, it is classified as an anomaly. As the optimal threshold is determined by maximizing F1 on Build C, the performance metrics reported in Table 4.1 reflect in-sample calibration performance on this build.

Ground truth labels for Build C were established by domain experts at Colibrium Additive based on visual inspection of the LayerQam imagery. Layers up to including 320 mm in build height were labelled nominal, while layers beyond this point were labelled anomalous, reflecting the observed onset of quality degradation. This resulted in 337 and 98 anomalous layers. These labels served exclusively as evaluation references and were not used during model training.

4

Results

This chapter presents the performance evaluation of the four model architectures investigated in this thesis: the Double Translation model (RM2L), the Autoencoder (AE), the Rake Translation (R2L), and the CAD translation model (M2L). Results are reported in terms of classification metrics, per-layer anomaly scores, and training behaviour.

4.1 Model Performance Comparison

Table 4.1 summaries the anomaly detection performance across all four architectures evaluated at their respective optimal classification thresholds. Figure 4.1 provides a visual comparison of the same metrics.

The RM2L model achieves the highest overall performance with an F1 score of 0.9846, a precision of 0.9897, and a recall of 0.9796. The AE model follows closely with an F1 score of 0.9648, matching RM2L in recall while exhibiting slightly lower precision. The single-input translation models perform considerably worse, R2L achieves an F1 score of 0.8526, and M2L records the lowest F1 score of 0.7647. Notably, M2L exhibits the highest recall of 0.9286 across the single-input models, at the cost of a substantially reduced precision of 0.6500, indicating a noticeable tendency towards false positives.

Table 4.1: Performance metrics for anomaly detection across the different model architectures. The RM2L (Double Translation) architecture yields the highest F1 score by fusing CAD and Rake data evaluated on the calibration build (Build C). Note that threshold was selected to maximize F1 on this same dataset.

Model Architecture	Opt. Threshold	Max F1 Score	Precision	Recall
RM2L (Double Translation)	0.5685	0.9846	0.9897	0.9796
AE (Autoencoder)	0.5301	0.9648	0.9505	0.9796
R2L (Rake Translation)	0.6782	0.8526	0.8804	0.8265
M2L (CAD Translation)	0.6476	0.7647	0.6500	0.9286

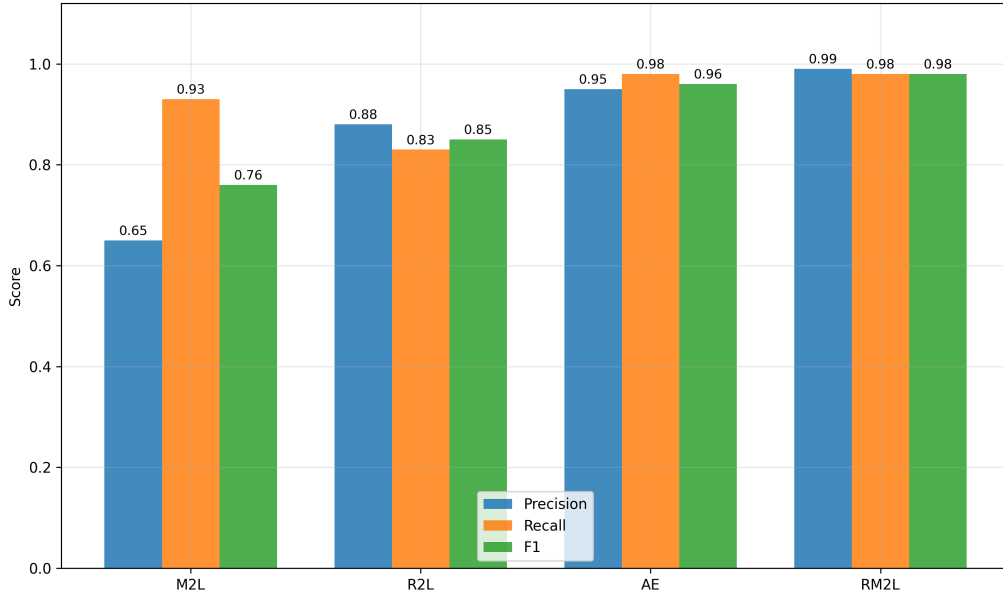


Figure 4.1: Visual comparison of Precision, Recall, and F1 scores across model architectures (see Table 4.1 for exact values).

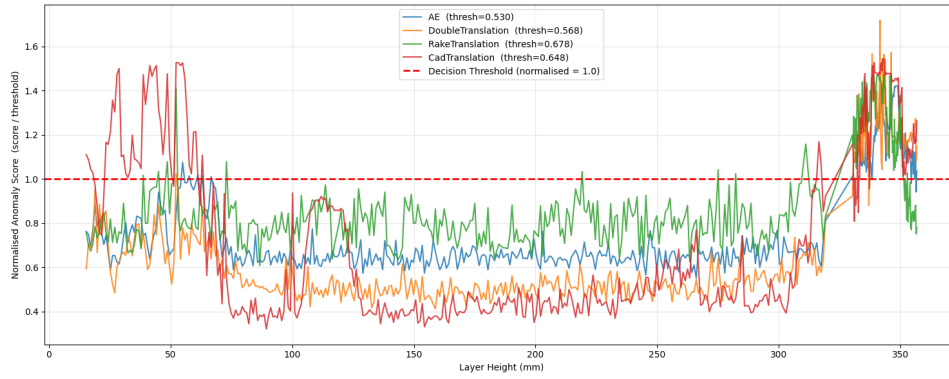
4.2 Anomaly Detection on Build C

Figure 4.2 presents a comparative analysis of anomaly detection across all models throughout the full build height of build C, alongside the turbine blade and corresponding cross-sectional area curve. Decision thresholds are normalized to a common scale allowing direct comparison across the different architectures. The RM2L model produces the most stable score classifications for the different nominal and anomalous layers. A noted pattern from all networks is elevated anomaly scores around layer 50 in the nominal region. This corresponds to a transition zone where the cumulative cross-sectional area increases significantly, as shown in Figure 4.2c and Figure 4.2b.

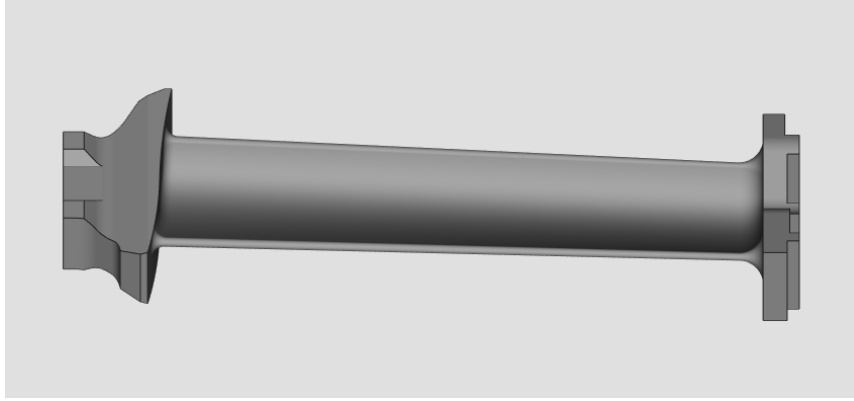
Figure 4.3 presents the detailed classification results for Build C using the best performing RM2L model. The confusion matrix (Figure 4.3a) shows that 96 out of 98 anomalous layers were correctly identified, and 336 out of 337 nominal layers were correctly classified, resulting in a single false alarm. The per-layer anomaly score curve (Figure 4.3b) shows that the reconstruction error remains consistently below the optimal threshold throughout the nominal build region, with exception for some layers in the beginning, and ultimately as the model enters the non-nominal region a clear and distinct spike is seen where the anomaly score exceeds the threshold for almost all layers.

4.3 Generalization Across Build Cycles

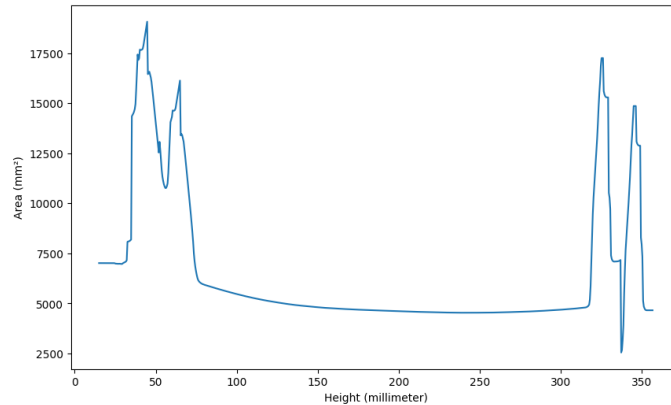
Figure 4.4 shows the anomaly score of the RM2L model across multiple build cycles, including both nominal builds and builds containing industrial anomalies. In



(a) Performance comparison of anomaly detection models on Build C. The x-axis represents build height (mm) and the y-axis the anomaly score (Equation 3.2). Decision thresholds are normalized to a common scale to allow direct comparison across architectures. RM2L (Double Translation) consistently shows the clearest separation between nominal and anomalous layers, with all layers beyond 320 mm being anomalous.

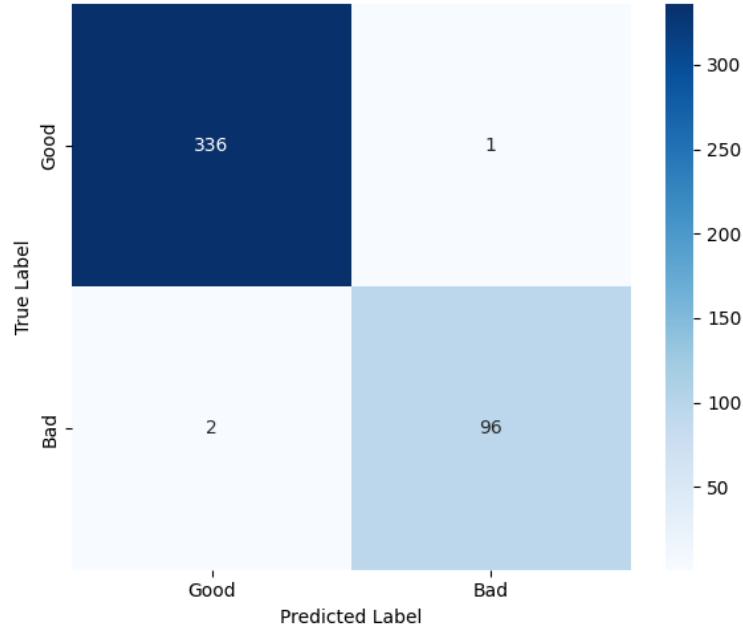


(b) Physical reference of the turbine blade, rotated 90 degrees to align with the vertical height axis of the plots. This orientation enables a direct visual correlation between detected anomaly spikes and the specific geometric features of the part.

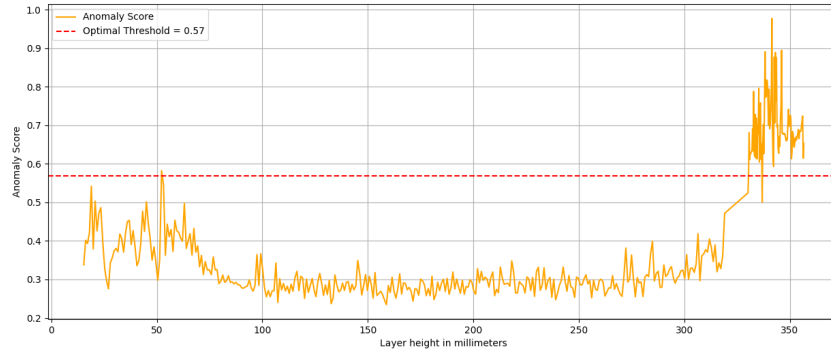


(c) Cross-sectional area measurements relative to the part height. This plot provides context for how geometric transitions and volumetric complexity correlate with the fluctuations in the anomaly scores observed in the detection models.

Figure 4.2: Integrated analysis of anomaly detection performance (top), physical geometry (middle), and volumetric cross-sectional area (bottom) throughout the total build height.



(a) Confusion matrix for Build C using Model RM2L.



(b) Anomaly score per layer height for Build C using Model RM2L with the optimal threshold as stated in Table 4.1.

Figure 4.3: Classification results for Build C using the best performing model, RM2L.

nominal builds, the anomaly score remains consistently below the decision threshold throughout the build height. In anomalous builds, exceedance of the threshold are observed at layers corresponding to known defect regions. This behaviour is consistent across the evaluated build cycles.

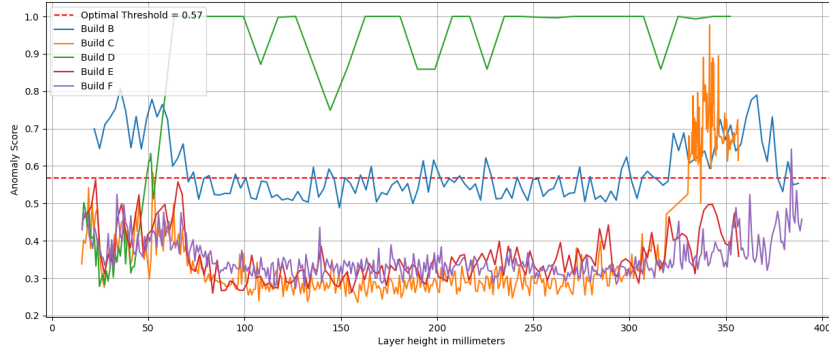


Figure 4.4: Comparison of anomaly scores for the RM2L (Double Translation) model across several different build cycles. The dataset includes both nominal builds and those containing various types of industrial anomalies, illustrating how the model’s reconstruction error fluctuates relative to the decision threshold depending on the specific build conditions and presence of defects.

4.4 Training and Validation Behaviour

Figures 4.5 and 4.5 present the training and validation loss and IoU curves for all four models.

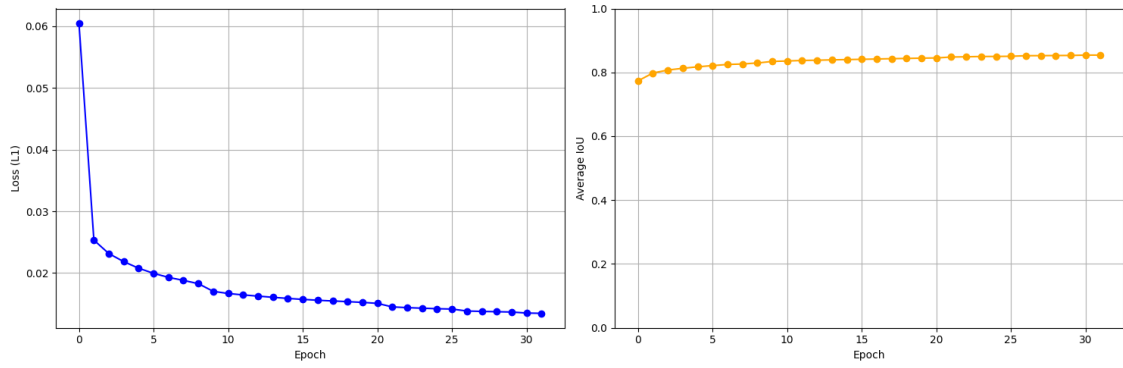
The RM2L model exhibits a smooth and steady decrease in training loss across all epochs, with the training IoU improving consistently throughout. The validation loss follows a general downward trend, though with considerable intermittent spikes, and the validation IoU improves gradually over the course of training. A persistent gap between training and validation IoU is observed throughout.

The AE model drops aggressively in training loss within the first epoch, after which it decreases slowly and steadily, with the training IoU improving consistently. The validation behaviour is noticeably unpredictable with large intermittent spikes in both loss and IoU, and no clear convergence trend.

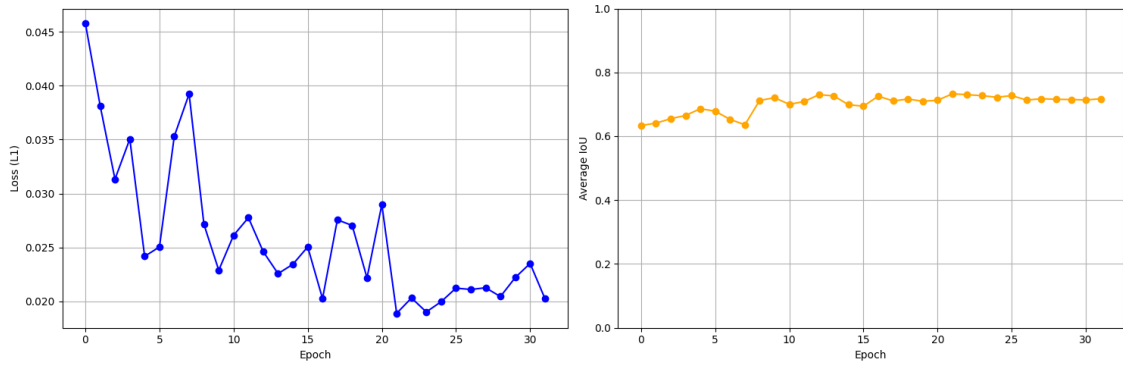
The M2L model trains smoothly over a greater number of epochs, with both training loss and IoU improving steadily. The validation loss reaches near its minimum within the first few epochs and subsequently oscillates within a narrow range, while the validation IoU remains stable. A gap between training and validation IoU persists across all epochs.

4. Results

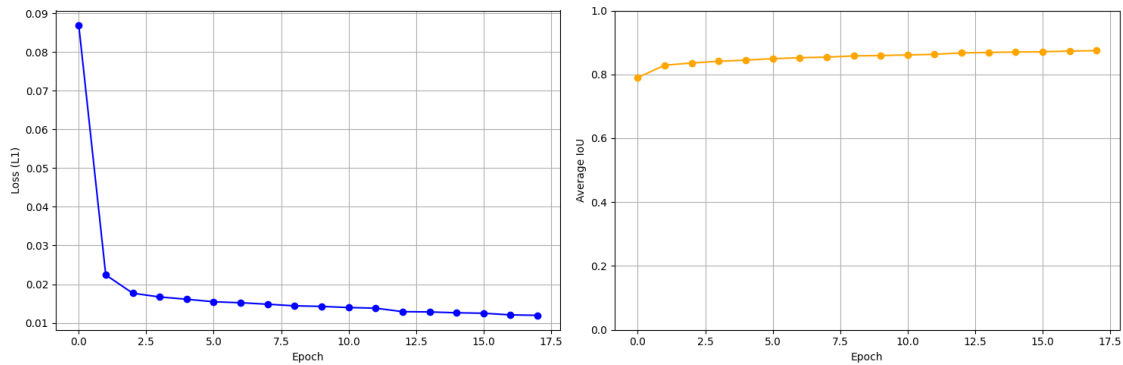
The R2L model exhibits a steadily decreasing training loss and consistently improving training IoU. The validation loss spikes sharply from the first epoch and remains substantially elevated throughout, with the validation IoU consistently and significantly below the training IoU. Training was consequently stopped early.



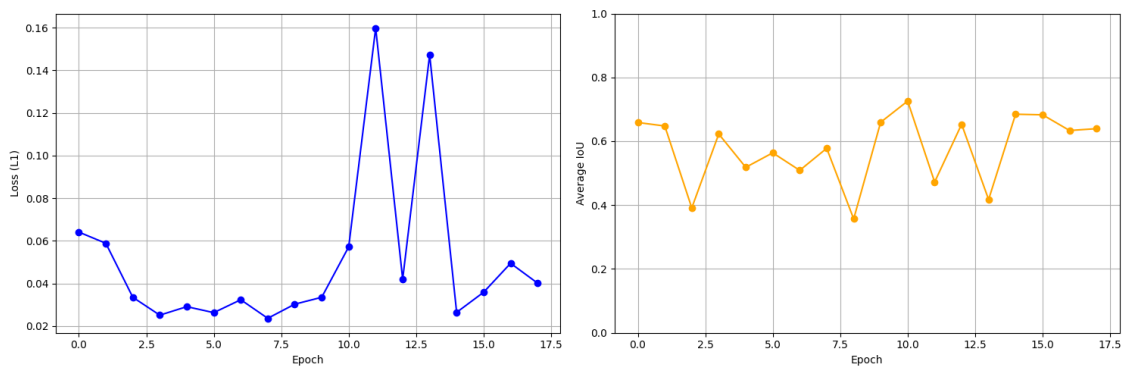
(a) Model RM2L - Training Loss & IoU



(b) Model RM2L - Validation Loss & IoU



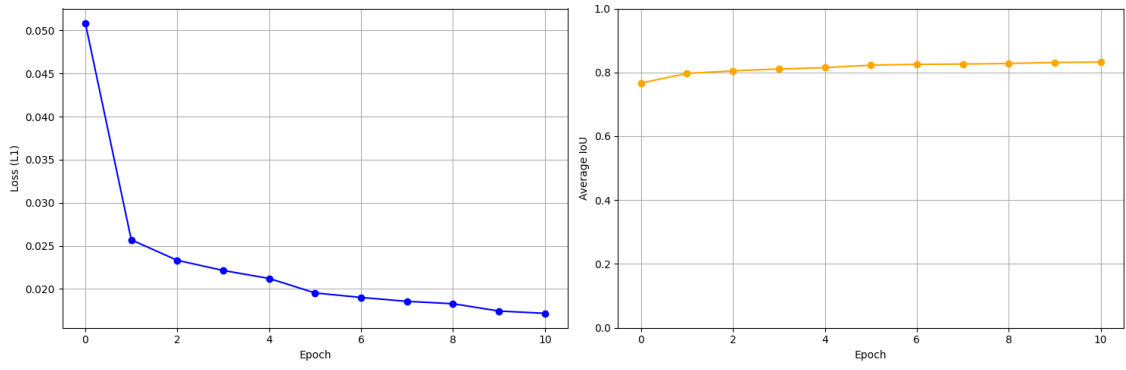
(c) Model AE - Training Loss & IoU



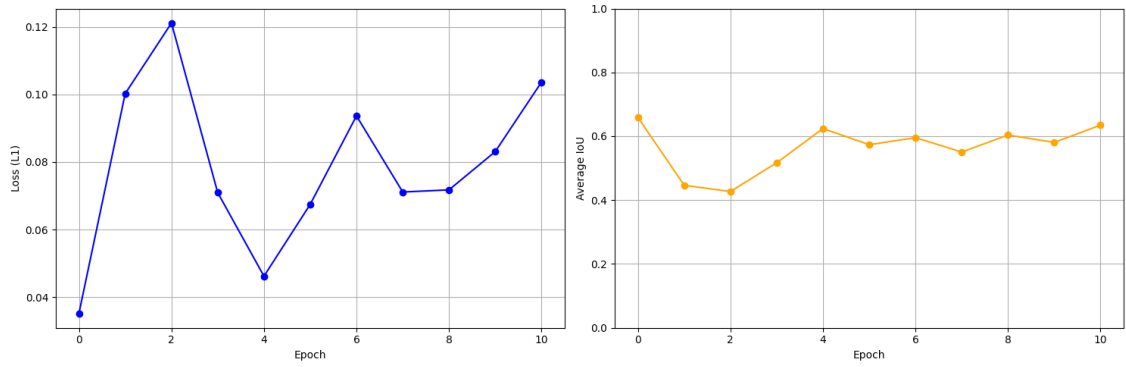
(d) Model AE - Validation Loss & IoU

Figure 4.5: Training and Validation Loss & IoU for all four models (1/2).

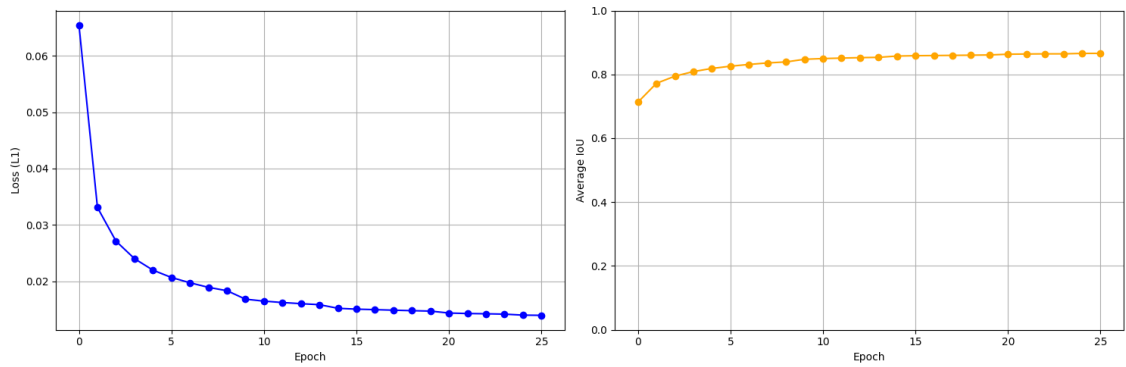
4. Results



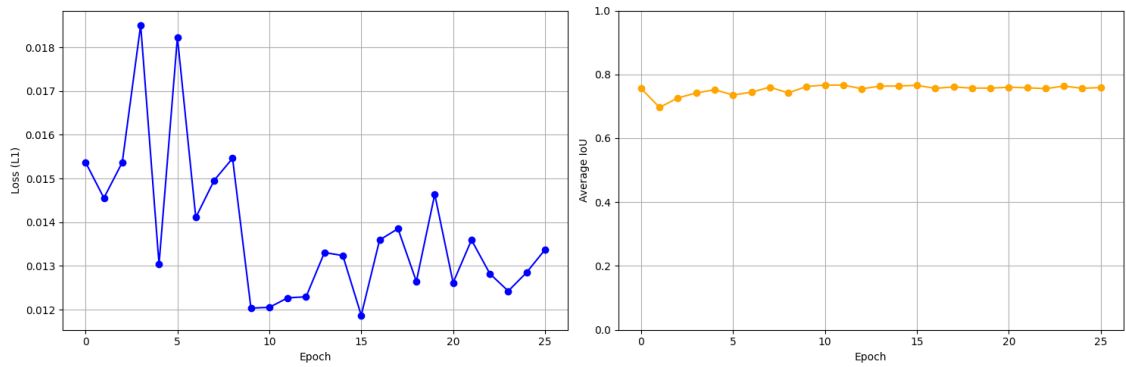
(e) Model R2L - Training Loss & IoU



(f) Model R2L - Validation Loss & IoU



(g) Model M2L - Training Loss & IoU



(h) Model M2L - Validation Loss & IoU

Figure 4.5: Training and Validation Loss & IoU for all four models (cont.).

5

Discussion

5.1 Combined Input and Autoencoder Performance

The RM2L model, which fuses both CAD and Rake inputs, achieves the best overall performance. The two input sources are complementary where the CAD mask provides unambiguous geometric guidance, while the Rake image supplies the physical context that prevents the model from over-penalising nominal process variation.

The AE model achieves a comparable F1 score, trailing RM2L by approximately 2 percentage points, with identical recall but slightly lower precision. This is a notable result given that the AE is a considerably more lightweight architecture that operates without any contextual input, relying solely on learning the distribution of nominal melt pool appearance. Despite the large spikes in the validation loss as observed in Figure 4.5d, the validation loss does not retain a sustained upward trend, indicating that the model has not collapsed as it finds itself back rather quickly. It can be believed to instead be due to the validation set containing enough variability to prevent stable convergence, however the learned nominal distribution is robust enough to support reliable anomaly scoring.

The similarity in performance between RM2L and AE raises the question of whether the marginal gain from the full I2I framework justifies its additional computational cost in a production deployment, where inference time is a practical constraint. It also reinforces the observation that the quality and nature of the input data are more decisive factors than architectural complexity alone.

5.2 Ablation Study: The Role of Input Data

While RM2L and AE demonstrate strong performance, the single-input translation models reveal the importance of the input context. As presented in Table 4.1, M2L records the lowest F1 score of 0.76, followed by R2L at 0.85, highlighting that neither geometric nor physical context alone is sufficient, compared to RM2L’s F1 score of 0.98.

The M2L model might be expected to perform well, as the CAD mask provides a geometrically perfect representation of the expected melt area, free of noise. However, this comes at the cost of all physical context. Minor variations in the process state, such as surrounding thermal intensity, powder bed inhomogeneities, or slight

parameter fluctuations, are entirely absent from the CAD input, yet present in the actual melt image. Since the model has no mechanism to account for these physical variations, any such deviation will produce a residual error that the model cannot distinguish from a true anomaly, resulting in the observed high false positive rate and low precision.

The training behaviour of M2L (Figures 4.5g and 4.5h) reinforces this interpretation. Notably, the validation loss remains close to the training loss throughout, which alone would imply the model is performing well. However, the persistent IoU gap suggest that the model has learned to reconstruct the training distribution without capturing the physical variation present in unseen builds.

The R2L model addresses this limitation by substituting the CAD mask with an actual Rake image. As the Rake image captures the physical powder bed prior to melting, the part geometry remains clearly distinguishable from the background while also providing physical context that the M2L model lacks. This dual contribution explains the approximately 9 percentage point improvement in F1 score over M2L. However, as observed in Figure 4.5e and 4.5f, R2L exhibits the most severe overfitting of all models, with training stopped very early due to immediate validation divergence. This suggests that while the Rake image provides richer input than the CAD mask, the model fails to generalise its translation to unseen physical conditions when trained on a single build. This raises the question of whether data augmentation or a more diverse training set would be sufficient to mitigate the overfitting behaviour observed in R2L.

Comparing M2L and R2L further, M2L surpasses R2L in recall, indicating greater sensitivity to anomalies, while R2L greatly outperforms M2L in precision, producing far fewer false alarms on nominal layers. This trade-off reflects the fundamental difference in their input contexts. M2L flags anything deviating from the geometrically ideal, while R2L leverages physical context to better distinguish true anomalies from nominal process variation.

5.3 Geometric Data Imbalance and Anomaly Score Spikes

A recurring observation across all models is the presence of elevated anomaly scores in the transitional regions of the build, visible in both the per-model comparison on Build C (Figure 4.2a) and across multiple build cycles (Figure 4.4). Each build contains eight turbine blades, four of which begin at an offset layer height relative to the other four. Examining the blade geometry (Figure 4.2b), notable wider regions are present at the lower and upper extremes, causing the total cross-sectional area to increase successively as the build transitions through these regions, as illustrated in Figure 4.2c. The greater surface area results in higher cumulative infrared intensity, and since these transitional regions constitute only a small fraction of the total build, they are underrepresented in the training distribution relative to the more uniform

mid-build layers. This directly explains the localised spikes in reconstruction error observed at these heights across all models.

5.4 Overfitting and Architectural Considerations

All models exhibit some degree of discrepancy between training and validation behaviour, reflecting both the inherent complexity of the thermal imaging data and the constraint of training on a single build. No hyperparameter optimisation was performed across any I2I models, as the study is designed as a controlled ablation where the network architecture is held constant to isolate the effect of input modality. It is worth noting that the AE operates under a somewhat different architectural configuration, with distinct layer parameters and a different batch size, as described in the Method chapter. The remaining training parameters are however kept consistent across all models. Future work could explore whether independently optimising the architecture for each input configuration yields further improvements, or whether the current configuration is already near the performance ceiling given the available training data.

5.5 Conclusion

The results demonstrate strong practical potential for replacing the current manual and heuristic-based inspection routines. Importantly, the CAD masks required by the best-performing RM2L model are generated prior to the build commencing and are therefore always available at inference time with no additional overhead, making the model immediately deployable within the existing manufacturing workflow. The strongest model (RM2L) requires no manually annotated data for training, directly addressing both the data confidentiality constraints and the time-consumption constraints identified in the problem description. Furthermore, as the models learn solely from nominal data, any deviation from expected behaviour is flagged regardless of its type, making the approach essentially capable of detecting novel or uncommon defects where labelled examples are scarce or unavailable. The consistency of the anomaly score behaviour across multiple build cycles (Figure 4.4) further suggests that the approach is scalable beyond a single build. While quantitative metrics are unavailable for these builds due to the absence of ground truth labels, qualitative evidence proves that the approach generalises beyond Build C. Future work should explore validation across additional geometries and process parameters, as well as independent hyperparameter optimisation for each input configuration, since the current study intentionally held the architecture constant to isolate the effect of input modality.

Bibliography

- [1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [2] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.
- [3] Ian Gibson, David Rosen, Brent Stucker, and Mahyar Khorasani. *Additive Manufacturing Technologies*. Springer, 3rd edition, 2021.
- [4] Davide Cannizzaro, Paolo Antonioni, Francesco Ponzio, Manuela Galati, Edoardo Patti, and Santa Di Cataldo. Machine learning-enabled real-time anomaly detection for electron beam powder bed fusion additive manufacturing. *Journal of Intelligent Manufacturing*, 36(3):2105–2119, March 2025.
- [5] Temilola Gbadamosi-Adeniyi, Trevor McDonald, Scott Ferguson, and Tim Horn. Enhancing in-situ monitoring in additive manufacturing with conv1d autoencoder: A deep learning approach to analyze total electron emission 1d signal. *IISE Annual Conference.Proceedings*, pages 1–7, 2024. Copyright - Copyright Institute of Industrial and Systems Engineers (IISE) 2024.
- [6] Léopold Le Roux, Chao Liu, Ze Ji, Pierre Kerfriden, Daniel Gage, Felix Feyer, Carolin Körner, and Samuel Bigot. Automatised quality assessment in additive layer manufacturing using layer-by-layer surface measurements and deep learning. *Procedia CIRP*, 99:342–347, 2021. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.
- [7] Marco Grasso and Bianca Maria Colosimo. A review of the current state-of-the-art on in situ monitoring in electron beam powder bed fusion. *Progress in Additive Manufacturing*, 9(5):1449–1466, October 2024.
- [8] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*, pages 1–5, 2018.
- [9] Rashmi Siddalingappa and Sekar Kanagaraj. Anomaly detection on medical images using autoencoder and convolutional neural network. *International Journal of Advanced Computer Science and Applications*, (7), 2021.
- [10] Additive manufacturing — general principles — terminology. Standard, International Organization for Standardization, nov 2021.
- [11] Ian Campbell, Olaf Diegel, Joseph Kowen, Noah Mostow, and Terry Wohlers. *Wohlers Report 2022: 3D Printing and Additive Manufacturing Global State of the Industry*. Wohlers Associates, Fort Collins, CO, 2022.
- [12] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2nd edition, 2003.

- [13] Jianhua Wang, Fanhuai Shi, Jing Zhang, and Yuncai Liu. A new calibration model of camera lens distortion. *Pattern Recognition*, 41(2):607–615, 2008.
- [14] Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.
- [15] Valerio Biscione and Jeffrey S. Bowers. Convolutional neural networks are not invariant to translation, but they can learn to be, 2021.
- [16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
- [17] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se



CHALMERS
UNIVERSITY OF TECHNOLOGY