Technische Universität München
TUM School of Engineering and Design

Anomaly Detection in Process Monitoring Data of Additive Manufacturing by Neural Networks

Quantifying Artificial Intelligence in Non-Destructive Testing

Jonas Holtmann

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. Katrin Wudy

Prüfende der Dissertation:

1. Prof. Dr.-Ing. Christian Große
2. Prof. Dr. techn. Peter Mayr

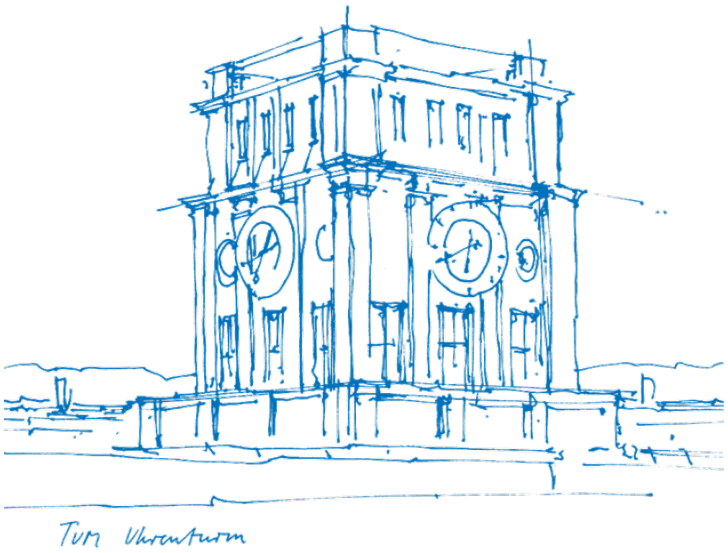
Die Dissertation wurde am 01.08.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 13.01.2025 angenommen.

Chair of Non-Destructive Testing
TUM School of Engineering and Design
Technical University of Munich

Anomaly Detection in Process Monitoring Data of Additive Manufacturing by Neural Networks

Quantifying Artificial Intelligence in Non-Destructive Testing

Jonas Holtmann



To everyone that supported me over the years. Thank you!

Abstract

Metal additive manufacturing (AM) offers the opportunity to produce geometrically optimized, lightweight structures for various use cases. In the aerospace industry, this potential comes at the cost of an increased testing effort, as additive manufactured parts must be inspected by non-destructive testing. The current inspection is performed by computed tomography (CT), which is time- and cost-intensive. In contrast, online monitoring (OM) of the printing process should allow for the evaluation of the printed part during manufacturing, saving cost and time and allowing for the manufacturing of larger and more complex parts that can currently not be inspected by computed tomography. Even though there are various monitoring systems available commercially and in academia, the current systems do not allow for the reliable detection of actual defects in the finished product as they focus solely on process deviations without a link to the resulting product. As a result, the process monitoring, particularly the data analysis, represents the current bottleneck in accelerating the AM industrialization process.

This thesis uses machine learning (i.e., neural networks) and a data correlation approach to solve this bottleneck. By correlating the online monitoring data with post-process testing data, i.e., computed tomography data, the effect of monitored anomalies on the finished sample is evaluated. In the first step, multiple convolutional neural networks (CNN) are trained to analyze the location and size of defects in laser powder bed fusion (LPBF) printed specimens based on the CT data. The performance of the CT CNNs is evaluated using a "qualified" dataset and use-case-specific metrics. It shows a high probability of detection with a low false alarm rate. In the second step, this data is used as a reference for training multiple CNNs based on the online monitoring data. The OM CNNs use the meltpool radiation measured during the printing process to detect defects in the printed part. Combined with the data fusion of two domains (post-processing CT and in-process monitoring), this supervised learning approach shows great potential for detecting pores in LPBF parts. The CNN results are quantified using custom-designed metrics and the probability of detection. Besides showing the general feasibility of the proposed system, the influence of the input monitoring data and the transferability to other sensor setups and defect types is investigated. Additional experiments are conducted to get a deeper insight into the performance and limitations of the CNN. The results indicate a well-performing and adaptable model that reliably detects defects based on monitoring data.

The described approach is independent of the sensor setup, and the conducted experiments indicate that it can be transferred to other printers and sensors. Therefore, the presented study provides a valuable step toward the reliable online monitoring of the LPBF process and, hence, towards solving the current technological bottleneck for the further adoption of LPBF in safety-critical industries.

Contents

1 Introduction

1.1 Motivation

The possibility of designing and producing parts of almost arbitrary complexity has led to a steep increase in additive manufacturing (AM) in different industries. Particularly in the aeronautic sector, AM's potential for producing lightweight, topologically optimized parts has been investigated for years. However, the design freedom offered by AM often results in complex part geometries, which represent a major challenge for the non-destructive testing (NDT) of such parts. In the case of metallic additive manufactured parts with a simple to medium complexity in geometry, this evaluation is performed by a cost- and time-intensive computed tomography (CT) scan (see Figure 1.1 top path). The CT scan can make up to 50% of the overall product cost. Besides the cost factor, the size and geometry of the current AM portfolio are limited by the capabilities of the CT machines, as the CT machine can only screen objects of limited size. This has hindered the further use of AM parts as safety-relevant components in aviation. Beyond the aerospace industry, the World Economic Forum has also identified the current costs per part and the quality control of AM parts as the two main bottlenecks for all industries using AM [Basso et al., 2022].

To expand the range of possible application cases of AM parts to safety-relevant structures and, at the same time, decrease the production and inspection cost, an online monitoring (OM) system of the manufacturing process would be advantageous. Online monitoring systems measure relevant quantities during the build process, allowing an in-process evaluation of the manufactured part. There exists a variety of different approaches and systems to monitor the AM process. They include systems provided by the machine suppliers, custom-designed systems, and research prototypes by universities and research facilities. As of today, these systems can capture a vast amount of data but are not able to provide reliable detection of anomalies in the finished part due to insufficient evaluation techniques (see Figure 1.1 bottom path). Instead, the current systems focus on detecting process deviations or anomalies. These deviations may or may not lead to a defect in the finished part. Therefore, while these systems provide valuable insight into the printing process, they cannot determine the existence of defects in the finished product. Hence, the in-process data analysis represents the current technological bottleneck for the industrial implementation of AM on a large scale.

This thesis aims to resolve this bottleneck using artificial intelligence algorithms, i.e., convolutional neural networks (CNNs). These CNNs will be trained to analyze OM data to detect and locate defects in the printed part. To achieve this, a digital twin of the AM part is generated. The digital twin contains the online monitoring data gathered during the printing process as well as the CT data of the corresponding finished part. The CT data allows for the precise detection and localization of defects in the finished part. By combining the CT and OM data, a link between characteristic features in the OM data and actual defects in the produced part can be established. This helps in bridging the gap between the observance of process deviations and the effect of those deviations on the actual part quality (see Figure 1.2). As a result, the proposed approach can distinguish between relevant deviations (which produce defects in the printed part) and irrelevant deviations, which do not influence the part quality (regarding the creation of pores).

To achieve this overarching goal, two additional research gaps have been identified. Firstly, the automatic analysis of CT data and, secondly, the qualification and interpretation of CNNs in NDT. The current qualified procedure for evaluating CT scans in the aerospace industry is a purely manual approach. An inspector visually inspects the entire scan and manually marks defects. This process is very time-consuming and responsible for the majority of the time and cost associated with the CT inspection. Hence, the development of an automatic and reliable labeling process for CT scans is a prerequisite for the evaluation of the data used in this study while at the same time representing a valuable research contribution on its own.

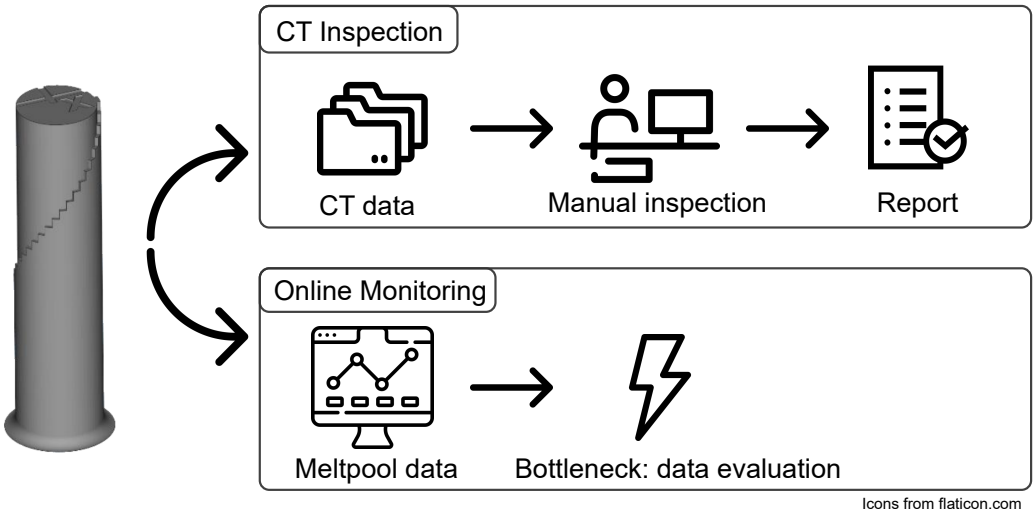


Figure 1.1 Current workflow for evaluating additively manufactured parts. Evaluation is performed by a human expert based on computed tomography scans (top path). The use of online monitoring data is currently limited by the analysis of such data (bottom path).

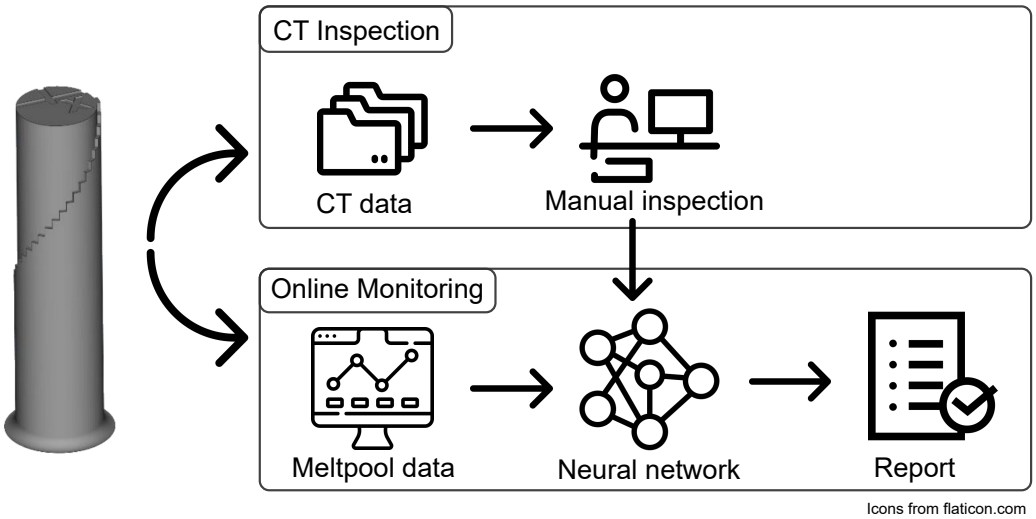


Figure 1.2 Basic idea introduced in this thesis: The expertise available in the CT domain is used to train a convolutional neural network to automatically analyze the online monitoring data.

Secondly, the qualification and interpretation of CNNs in aerospace, particularly in the NDT domain, is also an essential prerequisite for implementing the proposed approach later in the industry. While the qualification mainly depends on the regulatory framework by the authorities (i.e., the EASA), the interpretation and quantification of the CNN results is an active field of research in itself. Here, the definition and adoption of use-case-specific metrics for evaluating NDT CNNs offer the first steps to gain valuable insights into their performance, robustness, and transferability.

In summary, the presented approach aims to solve the two main bottlenecks for the further industrial usage of metal AM: quality assurance and costs. As pointed out, previous works and industrial systems focus on detecting process deviations rather than defects in the finished part. By developing and training a machine learning pipeline based on OM and CT data, this research (and industrial) gap is addressed. In doing so, this thesis also proposes the first steps for the additional research gaps of automatic CT analysis and the quantification and qualification of CNNs in NDT.

1.2 Focus of the Thesis

As described above, this thesis aims to contribute to the identified research gaps by combining the fields of non-destructive testing, additive manufacturing, and artificial intelligence. The overall goal is the development of an online monitoring system for the additive manufacturing process based on CNNs. This should facilitate the use of AM in safety-critical applications as it increases confidence in the part quality and, at the same time, drastically reduces the costs of manufacturing. To achieve this, a data science pipeline incorporating CNNs is developed (see Figure 4.1).

CNNs have demonstrated their potential in analyzing complex volumetric data in medical imaging. In contrast to conventional image analysis, CNNs do not require the definition of strict rules by an expert. Instead, a CNN extracts relevant information from the data by adapting to it. This process is called training and requires a large and diverse data basis. In medical imaging, small anomalies in large volumes have to be detected and segmented. An example is the segmentation of tumors in different organs based on CT scans. Commonly, the tumor covers only a small fraction of the entire volume, making its segmentation a challenging task for AI-based systems due to a high-class imbalance. In recent years, different network architectures have been proposed and investigated in the literature. Adapting these results to the task of analyzing online monitoring data provides a well-founded research basis for further development. In particular, the use and adaptation of the well-performing U-Net architecture to new use cases shows great potential.

The general concept for developing the data science pipeline is outlined in the following. Firstly, a statistically relevant number of samples is produced by laser powder bed fusion (LPBF). The specimens are printed with parameters known to provoke common anomalies, such as lack of fusion and keyhole. During the printing, the process is monitored by an online monitoring system that captures the meltpool radiation. The gathered raw data is then processed by a custom-designed pre-processing pipeline to retain a maximum degree of relevant information while drastically reducing the amount of data. The produced specimens are scanned with an industrial CT machine and evaluated concerning anomalies. For this evaluation, a CT CNN is trained and validated. The CT CNN is able to detect and segment pores in CT scans. These results represent the reference data for the subsequent analysis of the pre-processed online monitoring data. Registering the two datasets (CT and OM) to a common spatial reference frame makes a correlation of the two datasets possible. This allows for the correlation of deviations in the process (from the online monitoring) with the location and size of defects in the finished part (from the CT). Training a CNN on this combined dataset enables the CNN to extract relevant features from the online monitoring data, which indicate pores in the finished part.

In contrast to conventional analysis of the process stability, which only highlights deviations in the process, such features indicate the presence of actual defects. Therefore, the proposed system should be able to differentiate between irrelevant process deviations, which do not influence the part quality, and relevant features in the online monitoring data, which indicate a pore in the printed sample. The performance of the trained model is evaluated by analyzing the obtained results and the influences of different parameters on the CNN predictions. Furthermore, the first steps for the quantification and possible qualification of machine learning algorithms are introduced. For this, established statistical methods from the field of NDT are adapted to machine learning.

Therefore, in summary, the thesis combines the three introduced fields of research. Firstly, it aims to enhance the non-destructive testing of AM parts by automating the CT evaluation and implementing an online monitoring system. Secondly, in the field of AI, the thesis promotes the idea of data fusion from different sensors and domains. Such a data fusion approach allows for the combination of different datasets in order to produce a more meaningful analysis. Lastly, it intends to increase confidence in the AM process. To this extent, the monitoring and analysis of the meltpool signature represents its own relevant research topic and an important industrial use case.

1.3 Structure of the Thesis

In the following, the structure of the thesis is outlined. Firstly, this study's theoretical background (Chapter 2) is presented. As pointed out, the interdisciplinary research topic combines the fields of additive manufacturing, non-destructive testing (i.e., computed tomography), and artificial intelligence (i.e., convolutional neural networks). Hence, the necessary background for all three fields is presented with additional references provided for inclined readers.

Secondly, the state of the art (Chapter 3) is summarized. Here, particular focus is placed on the combination of AI, NDT, and AM. In particular, the current research in the field of artificial neural networks and LPBF monitoring systems is highlighted.

Thirdly, Chapter 4 explains the concepts, methods, and experiments implemented within this thesis. These can be divided roughly into the data pre-processing (Section 4.2 and Section 4.3) and the experimental design of the neural networks themselves (Section 4.5 and Section 4.6). As a first part, the data pre-processing includes the manufacturing of the specimens with reference defects. The second step consists of collecting and preparing CT and OM data. Subsequently, the necessary steps for the registration of the data are outlined. Section 4.4 introduces use-case-specific evaluation techniques and metrics to quantify the performance of the developed CNNs. Section 4.5 and Section 4.6 present a selection of CNN trainings for CT and OM data. The training and evaluation of the CT CNN follow an iterative approach. In the first iteration, a basic semi-automatic evaluation of the CT data is performed. This evaluation requires significant manual interaction but provides a sound baseline for the training of the first CT CNN. Based on this CT CNN, a more extensive and diverse dataset is evaluated and subsequently refined manually. This process is repeated numerous times to further enhance the CNN performance and data quality. A selection of the data is refined by qualified inspectors, generating a "qualified" label map, which is used as the reference for later evaluation. This iterative approach is described in Section 4.3.3 and Section 4.5 and allows for labeling large datasets with reasonable effort. The results of the CT CNN are used to train the OM CNN as they provide the location and size of defects in the printed part. In the scope of this thesis, a variety of different CNNs are trained. Only a selection of those is highlighted in Section 4.6. The presented experiments aim to show the general feasibility (Section 4.6.1 and Section 4.6.2), the influence of different sensors on the networks (Section 4.6.3) and the transferability and robustness of the CNNs (Section 4.6.4 and Section 4.6.5). In combination, the experiments not only show the potential of the proposed approach but also investigate its limitations and applicability to other use cases.

Fourthly, the results of the previously introduced experiments are presented in Chapter 5. The performance of the models is evaluated based on the custom metrics defined in Section 4.4. The evaluation of the experiments follows the line of Chapter 4 and analyses the conducted experiments with a focus on the corresponding targets stated in Chapter 4. As the experiments are conducted iteratively, previous experiment findings are included in subsequent experiments' design. This is particularly true for the CT CNNs and their use to generate (qualified) label maps to train subsequent CNNs (CT and OM).

In Chapter 6, the presented results are summarized and discussed. Section 6.1 focuses on the findings for the automatic CT evaluation. Section 6.2 discusses the OM CNN's feasibility, explainability, and transferability to automatically analyze the LPBF process's monitoring data. It compares the presented results and shows the potential of the investigated approach. The most important limitations of the approach are highlighted in Section 6.3.

Chapter 7 concludes the most relevant findings of the thesis and puts them into the context of the current research. In particular, it summarizes the contribution to the identified research gap and highlights the potential for industrial usage. This also includes the outlook on future work necessary to enhance the proposed approach.

2 Theory

The multidisciplinary research presented in this thesis combines the fields of Additive Manufacturing, Non-Destructive Testing, and Artificial Intelligence. Hence, the theoretical basis for those three areas will be presented, and the relevant interactions will be highlighted.

2.1 Additive Manufacturing

In general, additive manufacturing (AM) can be defined as a

> process of joining materials to make *parts* (2.6.1) from 3D model data, usually *layer* (2.3.10) upon layer, as opposed to subtractive manufacturing and formative manufacturing methodologies [British Standards Institution, 2015].

There exists a variety of manufacturing technologies that rely on this principle. This work will focus on Laser Powder Bed Fusion (LPBF), also known as selective laser melting (SLM) or laser-based powder bed fusion of metals (PBF-LB/M) [ISO, 07.2019]. The terms AM and LPBF are used as synonyms in this work. LPBF is defined as an

> additive manufacturing process in which thermal energy selectively fuses regions of a *powder bed* [British Standards Institution, 2015].

The basic principle of LPBF is illustrated in Figure 2.1. The machine consists of a metal powder reservoir (seen on the left), a build platform (seen in the middle), and a container that collects excessive powder (not shown in the illustration). The build platform is adjustable in height. The printing process can be divided into separate steps.

In the first step, the roller or recoater disperses a thin layer of powder from the reservoir onto the build platform. Then the laser is focused onto the build platform by an optical system (scanning mirror). By moving the laser over the build platform, defined areas of metal powder are melted. When the laser melts the powder at these points, the powder particles are fused together when cooling down. The particles do not only bond with particles in the same layer but also with powder or previously melted structures of the adjacent layers. By lowering the build platform and recoating it with new powder, the melting process can be repeated to form three-dimensional structures. Hence, the manufactured part is printed additively layer by layer. In the last step, the part can be extracted by removing the unmelted powder and separating it from the build platform. The advantage of this manufacturing approach is the almost unlimited geometrical design freedom, which allows for the production of complex and use-case-specific parts [Lachmayer and Lippert, 2017].

In the following, the physical and technological background necessary for the understanding of the AM process is presented. First of all, the basic LPBF process is explained. Here, particular focus is placed on the terminology and parameters used to create reference anomalies later in this work. In Section 2.1.2, additional parameters and the concept of energy density are introduced. Section 2.1.3 extends on these concepts and focuses on the meltpool characteristics as the meltpool represents the primary interaction between laser beam and material. Different studies have shown its influence on part quality [Bidare et al., 2018; Brailovski et al., 2020; Gordon et al., 2020]. Therefore, it is one of the most relevant quality indicators for the process stability and the focus of the monitoring system used in this study. Lastly, in Section 2.1.4, anomalies in the finished part as possible negative consequences of process instabilities are introduced.

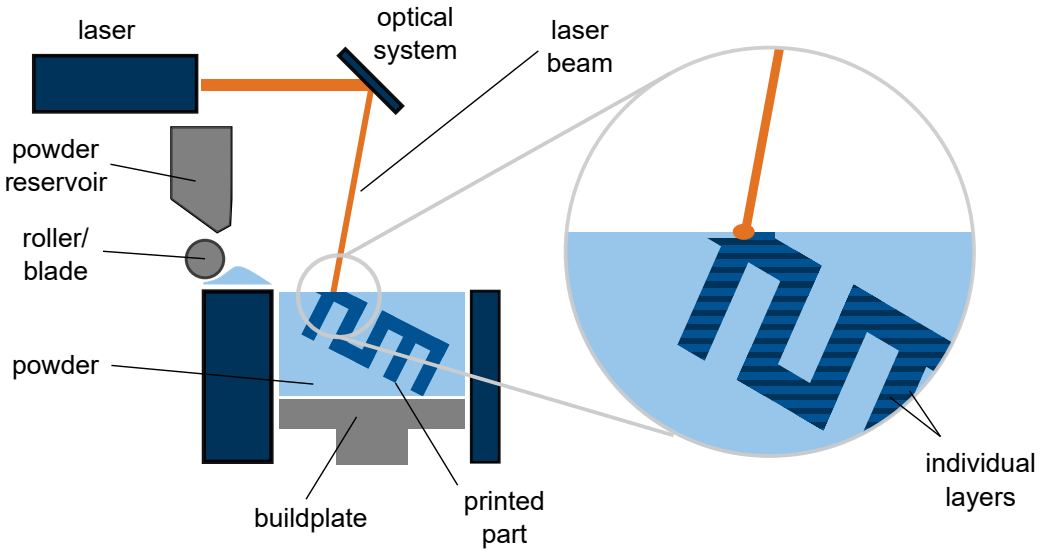


Figure 2.1 Schematic illustration of the Laser Powder Bed Fusion process (adapted from [Saracco, 18.09.2019]).

2.1.1 Process Overview

Different printing strategies (hatch strategies) have been developed to achieve high efficiency and reduce residual stresses. Figure 2.2 shows a schematic build plate with the laser track in black. Instead of lasering the specimens one by one, the entire build plate is divided into subsections (islands). Each subsection is lasered successively, reducing the thermal energy input to a specific area within a certain time span, hence reducing the risk of residual stresses [Meboldt and Klahn, 2018]. Within one island, the laser moves in meandering lines with a fixed distance. The distance between adjacent laser tracks is called **hatch distance**. At the end of each laser track, the laser turns in a circle to reach the starting point of the following laser track. This turning process influences the laser speed and, hence, the energy input per area. To avoid this process deviation, the laser is turned off during this time and is only switched back on once the laser is re-accelerated and realigned with the desired track. This control strategy is called **skywriting** and should ensure a stable process and well-defined laser tracks. The time in which the laser is switched off is referred to as skywriting lead time and is a configurable process parameter with considerable influence on the process stability [Duong et al., 2022]. At the border of two neighboring islands, the laser tracks are overlapped to avoid the creation of voids. In the following, this overlap is referred to as **hatch lines** as they are clearly visible at the borders of hatch patterns in the later presented meltpool data.

The scanning direction or hatch pattern is typically altered per layer to reduce the influence of the scanning direction on the properties of the finished part. A common strategy is the rotation of the pattern by 60 ° per layer.

It is critical that the hatch distance is adapted according to the meltpool size. If the meltpool is too small in relation to the hatch distance, regions between the tracks might not be melted or melted insufficiently. Besides other parameters, the meltpool size is significantly influenced by the laser focus. The focus of a laser is illustrated in Figure 2.3. The laser is focused using the optical system of the machine to achieve a high and constant energy density over the entire build platform. In the focal plane, the laser power is concentrated on the smallest area, resulting in the highest energy density. If the focal plane and the build plate do not coincide, the laser is defocused. Often the "defocussing" is referred to in terms of the distance of the focal plane to the build plate. The distance is positive if the focal plane lies above the build platform. If the focal plane lies below the build platform, it is negative [Ladewig, 2019]. Besides the distance of the focal plane from the build platform, the energy density depends on the geometry of the laser beam, which can be described by the Rayleigh length. The Rayleigh length describes the length along the laser beam at which the laser cross section has doubled in size [Hügel and Graf, 2009]. At this point, the laser energy is focused on double the area as in the focal plane, resulting in half the energy density. The focus of the

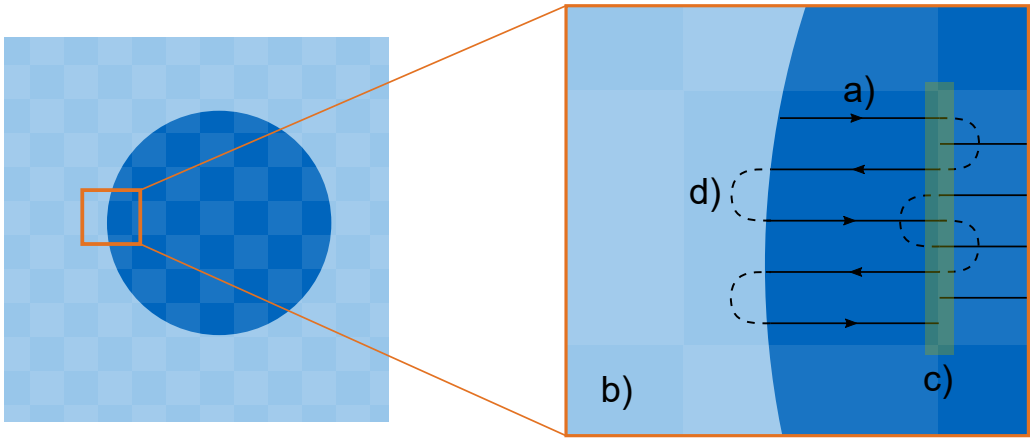


Figure 2.2 Schematic build plate with a) individual laser tracks, b) hatch pattern, c) hatch line, and d) illustrated skywriting tracks.

laser depends on multiple complex parameters (e.g., the temperature of the optical system) and, at the same time, is critical for a stable manufacturing process. Therefore, it has been identified as a relevant process parameter for the creation of anomalies and will be used in this study to provoke the formation of pores (Section 4.1).

Besides the laser, the powder and the shielding gas flow have been identified to influence the process stability. The entire printing process, particularly the melting of the powder, takes place in an inert gas environment. The inert gas prevents the oxidation of the metal powder. At the same time, the gas flow acts as a cleaning process by removing fumes and other by-products from the meltpool region. An unstable or insufficient gas flow may influence the part quality negatively, while an excessive gas flow might remove powder and hinder the correct building of the part. Due to the complex formation of the gas flow within the build chamber, the influence of the gas on the build process has to be regarded as locally resolved, which makes global monitoring of this process parameter insufficient for predicting possible local defects. [Ferrar et al., 2012; Ladewig et al., 2016]

The same holds true for the distribution of powder on the build plate. The printing process depends on a uniform powder bed as the process is optimized for a given layer thickness. The layer thickness has to be regarded as potentially varying over the build plate. Therefore, it cannot be monitored globally but requires a spatially resolved monitoring system. Possible deviations include a vibrating recoater blade, a damaged recoater, or an insufficient amount of powder [Fischer et al., 2022].

2.1.2 Process Parameters

Besides the above-mentioned global parameters, such as gas flow and hatch strategy, over 100 process parameters that influence the LPBF process have been studied [Oliveira et al., 2020]. Those parameters are determined empirically by the machine manufacturer and optimized to produce dense parts and minimize defects. In the following, the most relevant parameters identified in literature will be discussed [Spears and Gold, 2016; Letenneur et al., 2019; Vilanova et al., 2020]. For a more detailed summary, the reader is referred to [Spears and Gold, 2016; Oliveira et al., 2020].

Firstly, the **laser power** and the **scan speed** with which the laser travels on the build platform were identified as critical for properly fusing the metal powder. Figure 2.4 shows a qualitative process window defined by the laser power and scan speed. A region of optimal fusion exists where a well-tuned process should operate. If the laser power is too high relative to the scan speed, the process is prone to create keyholes due to excessive energy input. On the other hand, insufficient laser power may lead to a lack of fusion in the manufactured part as the energy input is not sufficient to melt the powder. In the case of a very high laser power combined with a high scan speed, the process becomes increasingly unstable, which might lead to balling [Oliveira et al., 2020].

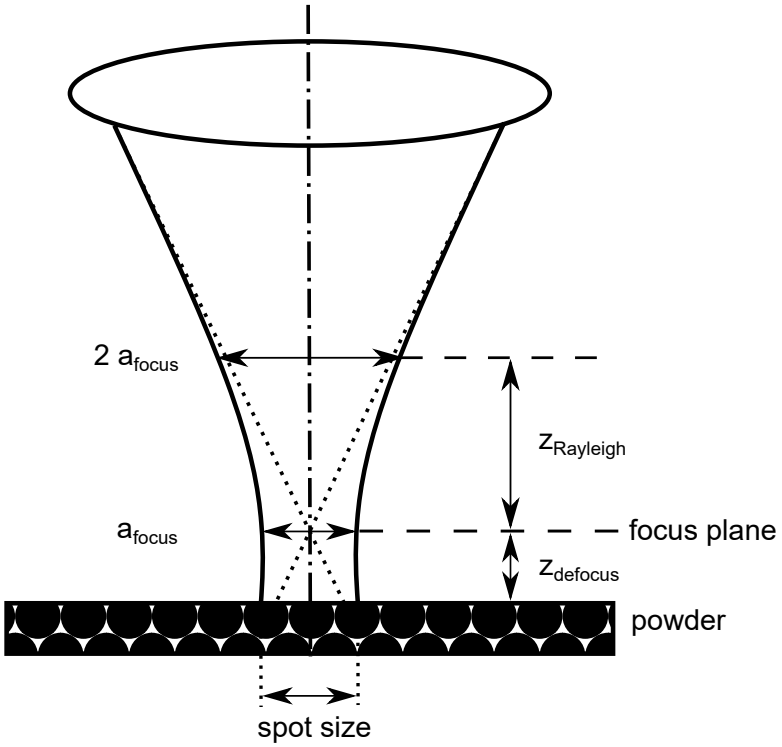


Figure 2.3 Illustration of the laser focus in relation to the build plate. The laser is focused in the focal plane and defocuses above and below this plane (adapted from [Hügel and Graf, 2009] and [Ladewig, 2019]).

Secondly, the **scanning strategy** and particularly the hatch distance have been identified as other relevant process parameters [Letenneur et al., 2019]. As pointed out in Section 2.1.1, if the distance between adjacent laser tracks is significantly too wide, the laser might not be able to melt all powder particles, resulting in an incomplete part. However, even if all powder particles are melted, the energy input might not be sufficient, leading to a lack of fusion between the laser tracks. If the hatch distance is chosen too small, already fused material is constantly remelted, which may lead to increased evaporation of specific alloys. Furthermore, the process is less efficient as more time and energy are needed to print a specific area [Dong et al., 2018; Oliveira et al., 2020].

Thirdly, the **layer thickness** considerably influences the part quality. The layer thickness is mostly influenced by (but not only dependent on) the height at which the build plate is lowered per layer. In this thesis and most other studies, the other influencing factors (e.g., imperfections due to the recoater) are neglected, and the build plate lowering is used as a defining parameter for the layer thickness. Similar to the hatch distance, an excessive layer thickness may lead to insufficient powder melting between layers. This can result in lack of fusion or even unmelted powder particles and ultimately in a lower density of the finished part [Letenneur et al., 2019].

Literature often combines the four introduced parameters to form the **energy density**. Where E_d is the energy density, P_L is the effective laser power, v is the scan speed, h the hatch distance and t is the layer thickness [Calignano, 2020]:

$$E_d = \frac{P_L}{v * h * t} \tag{2.1}$$

Different studies have shown a relationship between the energy density and the manufactured part's density. The optimal energy density and parameter sets have to be determined empirically or by simulation and largely influence the meltpool characteristics [Calignano, 2020; Brailovski et al., 2020].

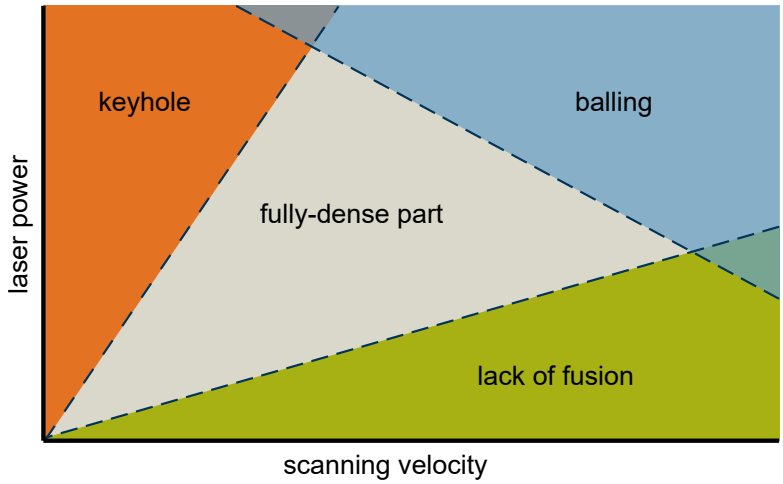


Figure 2.4 Qualitative process window for the LPBF process with the optimal process window in the middle section. The shown boundaries cannot be regarded as clearly defined but illustrate a wider zone of uncertainty in which the process is not clearly defined (adapted from [Oliveira et al., 2020]).

2.1.3 Meltpool Characteristics

The meltpool represents a primary feature in the laser-material interaction in the fusion process [Berumen et al., 2010; Craeghs et al., 2010, 2012; Doubenskaia, 2012; Doubenskaia et al., 2016; Clijsters et al., 2014]. The meltpool stability, dimension, and dynamic strongly affect the quality and stability of the LPBF process and, hence, the finished product [Thijs et al., 2010; Grasso and Colosimo, 2017]. Four main quantities characterizing the meltpool have been described: the size, the shape, the temperature intensity, and the temperature profile [Grasso and Colosimo, 2017]. They are primarily affected by the chosen process parameters and the beam-material interaction, which in turn influence the creation of most defect types (e.g., keyhole, lack of fusion). Additionally, the meltpool determines the geometrical accuracy of the finished part, its surface, and internal structures. Sub-optimal meltpool properties can also result in residual stresses, cracking, and lack of fusion [Mercelis and Kruth, 2006; Grasso and Colosimo, 2017; Lu et al., 2021]. The complex melting and solidification process is described in more detail in the following.

In general, the laser can interact in three different ways with material. It can be absorbed, reflected, or transmitted. In the case of metal powder, the energy is mainly absorbed or reflected. Only the absorbed energy of the beam is transformed into thermal energy, while the reflected energy does not contribute to the heating of the material directly. The amount of absorbed energy depends on multiple factors such as the material, the geometric and chemical surface texture, the angle of incidence, the temperature of the material, and the wavelength and polarisation of the laser [Meiners, 1999; Wagner, 2003; Hügel and Graf, 2009; Zeng et al., 2012; Ladewig, 2019; Li et al., 2021].

In the case of LPBF, two distinct interaction states exist: first, when the laser hits the powder bed before any powder is melted, and second, when the powder is melted and a meltpool already exists. In the first case, the laser is reflected and scattered multiple times between particles, transferring most of its energy into thermal energy in the powder particles. This state only exists briefly as the powder is melted quickly, creating a meltpool. Once the meltpool is created, the majority of the absorbed laser power is captured via the melt or melt surface [Rombouts et al., 2009; Ladewig, 2019].

Two common welding modes are differentiated in LPBF depending on the energy density. In conduction mode, a shallow meltpool without significant vapor creation is formed (Figure 2.5). The melt absorbs most of the laser energy, and the resulting thermal energy is transferred to the material. The geometry and depth of the meltpool are defined mostly by the thermal conduction of the surrounding material. Typically, the resulting welding track has an aspect ratio of width to depth below one. Conduction welding is the more stable process that creates little to no vapor, drastically reducing the probability of pores [Lane et al., 2020; Poprawe, 2005].

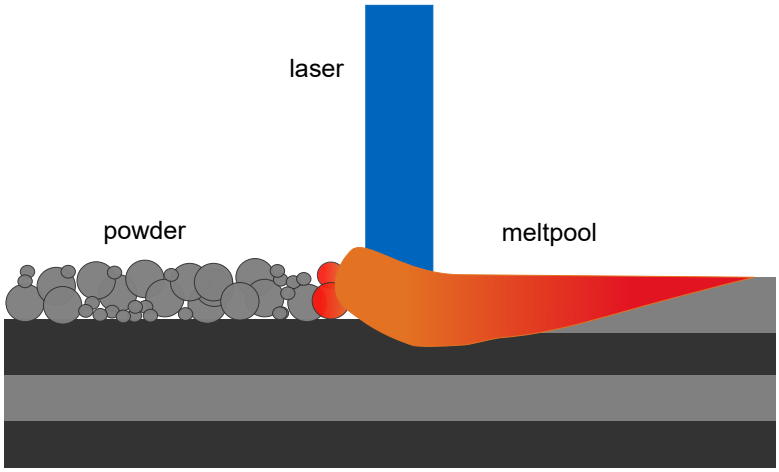


Figure 2.5 Schematic of the conduction welding process with a shallow meltpool and no vapor cavity (adapted from [Poprawe, 2005]).

In contrast, the keyhole mode is characterized by a significantly deeper meltpool (aspect ratio of width to depth above 1) [Poprawe, 2005]. Keyhole mode is reached by a higher energy density, which results in the heating of the metal above the boiling point, creating a vapor cavity in the meltpool (Figure 2.6) [King et al., 2014]. By multiple laser beam reflections inside the vapor cavity, more energy is absorbed than in conduction mode, increasing the energy input into the material significantly. The increased energy input and deeper meltpool lead to higher productivity and efficiency of the process. Additionally, the deep penetration of the meltpool allows for the fusion and possible remelting of lower layers, which may result in the healing of previously induced defects. On the other hand, the increased process dynamic and vapor creation may result in higher porosity and defects if the process is not controlled well [Quintino and Assunção, 2013; King et al., 2014]. The transition from a stable to an unstable keyhole (resulting in keyhole porosity) is subtle, and the physical phenomena leading to it still need to be fully understood [Huang et al., 2022]. Nevertheless, due to the higher productivity and the ability to process highly reflective metals, most LPBF processes operate in the keyhole welding mode. However, a dynamic transition between the modes is possible and is sometimes referred to as transition mode [Quintino and Assunção, 2013; Huang et al., 2022].

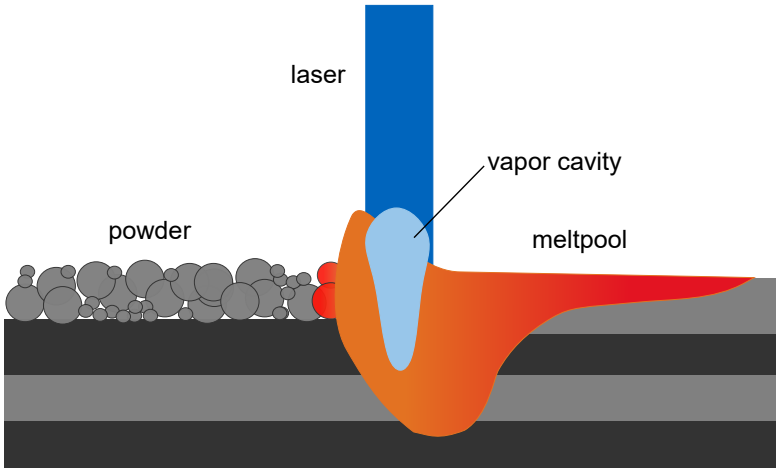


Figure 2.6 Schematic of the keyhole welding process with a deep meltpool and a vapor cavity (adapted from [Poprawe, 2005]).

2.1.4 Anomalies

The previous chapters introduced process parameters and their influence on process stability. In the following, anomalies as possible negative outcomes of such process deviations will be investigated in more detail. It is important to stress the difference between process deviations and actual anomalies in the finished part. While many studies focus on detecting process deviations, the relevance of such deviations often remains unclear. In particular, their consequence on the quality of the finished part is ambiguous as a process deviation does not necessarily result in an anomaly. Hence, when only focusing on process discrepancies without additional knowledge from post-process inspection (e.g., CT), an evaluation of the part quality is limited. Therefore, the presented correlation with CT represents an important addition as it links process deviation with actual anomalies, which will be introduced in the following.

Various types of anomalies have been described in research, such as keyhole pores, lack of fusion, inclusions, and cracks due to residual stresses [Harrison et al., 2015; Casati et al., 2016]. For a comprehensive list of anomalies and their possible causes, the reader is referred to the literature [Everton et al., 2016; Spears and Gold, 2016; Grasso and Colosimo, 2017; Zhang et al., 2017; Mahmoud et al., 2021]. This thesis focuses on keyholes and lack of fusion as they are identified as the most relevant anomalies within the project. Both lack of fusion and keyholes result in a higher porosity of the material. This is critical to the manufactured part as it strongly affects the crack growth and fatigue performance of the manufactured part [Moon et al., 2021; Khorasani et al., 2019].

Lack of Fusion

Lack of fusion is defined as the lack of union of powder and previously melted powder or substrate [DIN, 2007]. In most cases, they result from an insufficient energy density as the introduced energy is not high enough to melt the powder fully. Lack of fusion can be roughly divided into two classes: interlayer and intertrack lack of fusion. In the first case, the union of adjacent layers is not sufficient. This can be due to an excessive layer thickness or low energy input. In the second case, the fusion between individual laser tracks is insufficient. The most apparent reason is an excessive hatch distance, but other interference, such as geometrical inaccuracy in the scanner trajectory, can also lead to an intertrack lack of fusion [Zhang et al., 2017].

Other reasons for lack of fusion are the so-called "balling effect" and randomly occurring process deviations, such as turbulences in the shielding gas flow or the interference of the laser beam by foreign particles. Therefore, the creation of lack of fusion has to be regarded as a random occurrence within the printing process. Such anomaly might not be detected when only monitoring the print job globally (e.g., by an accompanying specimen). Instead, a local analysis of the printing process and its deviations is required [Ladewig, 2019].

Both classes of lack of fusion result in an inferior part quality with lower mechanical properties as they act as crack initiation points and significantly reduce the fatigue strength [Zhang et al., 2017; Hu et al., 2020].

Keyhole Porosity

Besides lack of fusion, keyhole porosity has been identified as the most prominent anomaly in the scope of this project. The creation of keyhole pores is closely related to the keyhole melting mode. As described in Section 2.1.3, the AM printing process is usually conducted in keyhole mode due to its higher efficiency. In this welding mode, a gas cavity is formed by the laser. In case of a transient or unstable melting mode, this gas cavity might collapse, trapping gas inside the material [Huang et al., 2022] (see Figure 2.8). The trapped gas forms spherical pores if it cannot escape due to the fast solidification process (see Figure 2.9) [Gong et al., 2014]. For a more detailed investigation of the different pore formation mechanisms, the reader is referred to the literature [Gong et al., 2014; Mancisidor et al., 2016; Martin et al., 2019a,b; Huang et al., 2022]. Like lack of fusion, keyhole pores can act as crack initiation points [Hu et al., 2020].

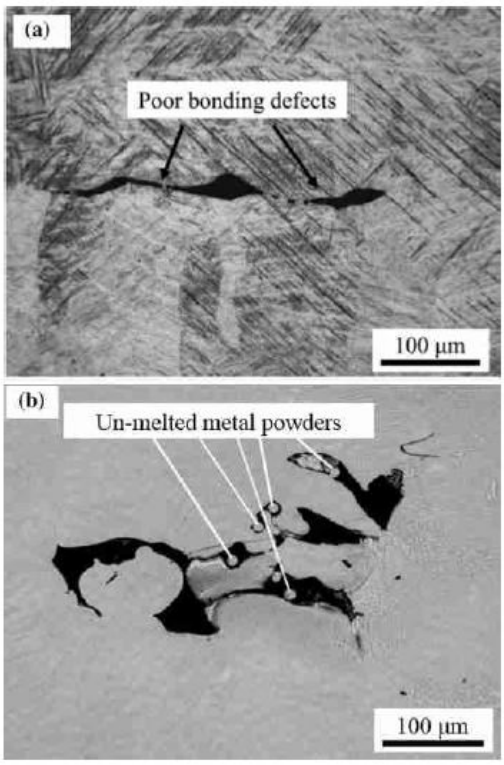


Figure 2.7 Microsection of a lack of fusion with: a) poor bonding b) poor bonding and unmelted powder within [Zhang et al., 2017; Liu et al., 2014].

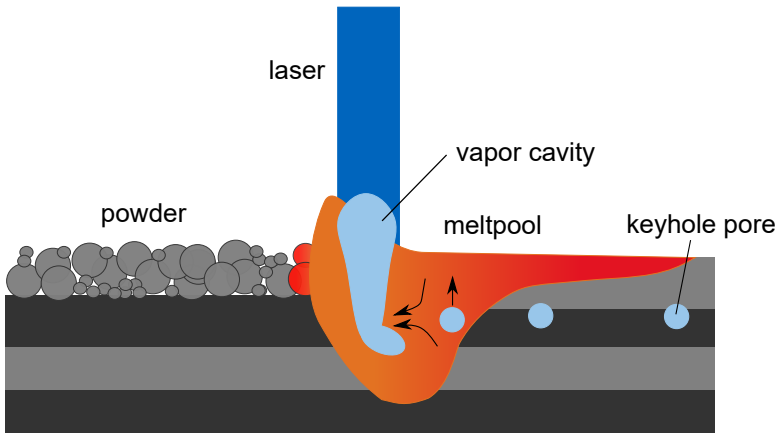


Figure 2.8 Keyhole creation due to an unstable gas cavity and gas trapping in the fast solidifying meltpool (adapted from [Ladewig, 2019]).

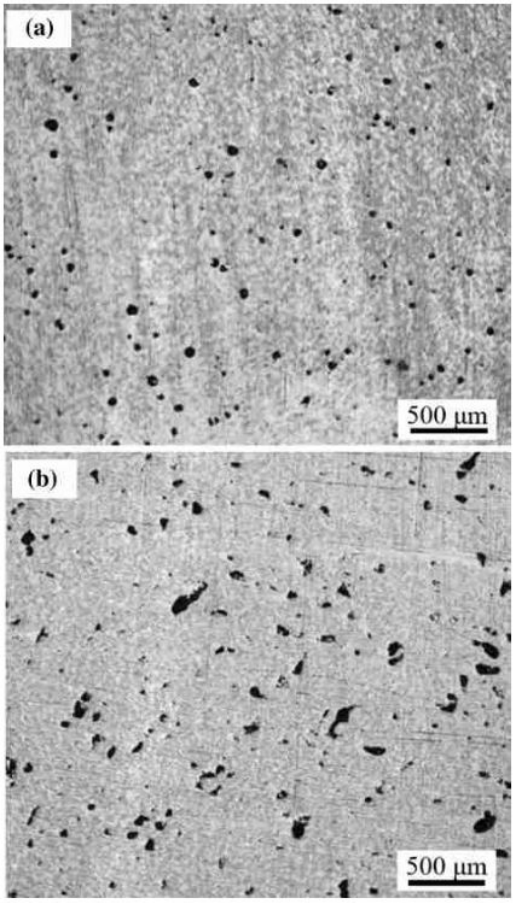


Figure 2.9 Microsection of a) spherical keyhole porosity b) lack of fusion [Zhang et al., 2017; Gong et al., 2015].

2.2 Computed Tomography

Computed tomography is the current standard for inspecting additive manufactured metal parts in aerospace. X-ray computed tomography is a radiographic NDT method that uses X-ray radiation to create a volumetric, internal image of a specimen. The physical background and working principles of computed tomography will be introduced in the following. Firstly, the physical fundamentals for creating and detecting X-rays will be introduced. Subsequently, the interaction of X-rays with the specimen matter will be discussed, and the reconstruction of a 3D volume from the gathered data will be explained. Lastly, artifacts and other interferences on the CT measurement are considered.

2.2.1 X-Ray Fundamentals

X-rays are electromagnetic waves with a high penetration power. While visible light only penetrates transparent materials, X-rays can pass through matter opaque to the human eye. Therefore, it is well suited to detect volumetric inhomogeneities (inclusions or pores) within a specimen [Schiebold, 2015b].

X-rays have a wavelength of between $0.01\,\mathrm{nm}$ and $10\,\mathrm{nm}$. According to Planck's energy-frequency relation, the energy E of the radiation increases with increasing frequency f, where h is the Planck constant, c is the speed of light, and λ is the wavelength.

$$E = fh = \frac{hc}{\lambda} \tag{2.2}$$

In reference to their energy level, X-rays with a longer wavelength ($> 0.1\,\mathrm{nm}$) are referred to as soft X-rays, whereas shorter wavelengths are termed hard X-ray [Bohm et al., 2019; Carmignato et al., 2018].

The most common source of X-rays for non-destructive testing is an X-ray tube. Figure 2.10 illustrates the general setup. The two main components of an X-ray tube are the cathode and the anode, which are placed inside a vacuum tube. The cathode is heated by an electric current (heating current), which leads to the thermionic emission of electrons. Those free electrons are accelerated toward the anode (target) by a high-voltage field between the cathode and anode. The vacuum allows for an acceleration of the electrons with minimal interference, e.g., by gas molecules. When the electrons hit the anode, they are decelerated abruptly. Around 99% of the kinetic energy of the electrons is converted into heat while around 1% is transformed into X-rays [Schiebold, 2015b]. Two possible interactions of electrons with the anode material lead to the generation of X-rays.

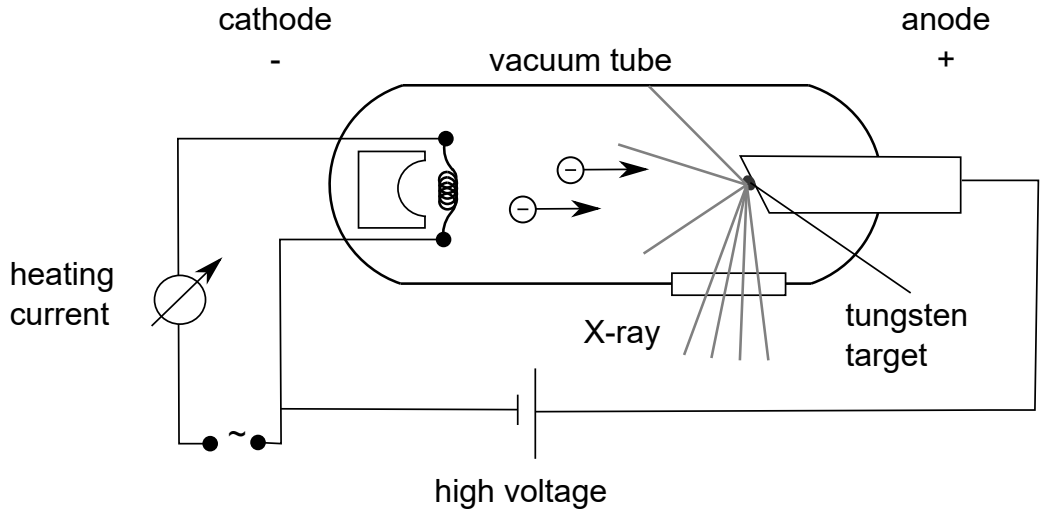


Figure 2.10 General setup of an X-ray tube with the cathode and anode in a vacuum tube (adapted from [Schiebold, 2015b]).

Firstly, the deceleration and deflection of electrons by the atomic nucleus (see Figure 2.11). The nucleus has a positive charge and, therefore, interacts with the negatively charged electron. The electron's kinetic

energy is transferred to the nucleus and instantly emitted as radiation. Therefore, the maximal radiation energy is reached when the total kinetic energy of the electron is transferred to the nucleus, and the electron is entirely stopped by directly hitting the nucleus. In the case of only partial electron deceleration, the radiation energy is only a fraction of the kinetic energy. Hence, the emitted radiation shows a spectrum of different energy levels. This radiation is called "Bremsstrahlung" (braking radiation). [Schiebold, 2015b]

Secondly, the accelerated electrons can interact with the bound electrons of the target material (=anode). When a free electron hits an inner-shell electron of the anode's atomic structure, this electron is knocked out of its shell, leaving a vacant spot in the electron shell of the atom. This vacancy is filled with an outer shell electron. As the electron moves to a lower energy level, the energy difference is emitted as radiation (see Figure 2.11). The energy levels that are more likely subject to electron transition are material-specific. Hence, the radiation emitted is characteristic of the material and is called characteristic radiation [Carmignato et al., 2018].

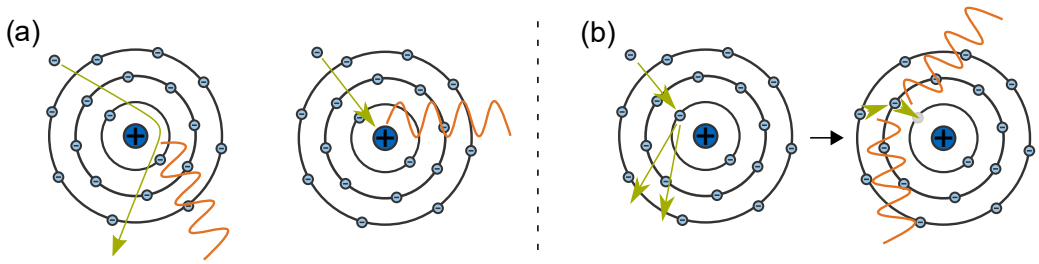


Figure 2.11 Principle of X-ray generation a) Bremsstrahlung: the accelerated electron is decelerated/ deflected by the atomic nucleus, which emits the transferred energy as X-ray radiation b) characteristic radiation: the accelerated electron knocks an electron out of its atomic shell, leaving a vacancy which is filled by a higher shell electron. The energy difference between the higher and the lower shell is emitted as X-ray radiation (adapted from [Carmignato et al., 2018]).

The energy of the X-ray radiation depends highly on the kinetic energy of the accelerated electrons, which in turn depends on the acceleration voltage between the cathode and anode. Therefore, the energy of the X-rays can be controlled by the voltage between the cathode and anode. A higher acceleration voltage leads to higher energetic X-ray radiation, allowing for higher penetration power. The voltage for industrial CT applications normally lies in the range between 40 kV and 600 kV [Kiefel, 2017]. In addition to the acceleration voltage, the X-ray spectrum is influenced by the heating current, the target material, and possible filters. A higher heating current leads to higher thermionic emission of electrons and hence a higher X-ray intensity at the cost of a possibly lower beam quality [Carmignato et al., 2018]. On the other hand, filters can be used to improve the beam quality of the X-ray. The unfiltered X-ray spectrum is composed of many different energy levels. The lower energy levels might influence the detection signal negatively and can be filtered by a physical filter placed between the X-ray tube and the specimen. Common filter materials are copper, aluminum, or tin [Carmignato et al., 2018; Schiebold, 2015b; Buzug, 2008].

2.2.2 X-Ray Imaging and Data Reconstruction

The generated X-ray beam is focused onto a specimen. When the beam passes through the specimen, the X-ray intensity is reduced exponentially. This process is called attenuation. The primary specimen properties influencing the attenuation are illustrated in Figure 2.12. The total attenuation is described by the Beer-Lambert law and the attenuation coefficient μ [Hertel and Schulz, 2017].

$$I(x) = I_0 e^{-\mu x} \tag{2.3}$$

Where $I(x)$ is the radiation intensity after the interaction with matter, I_0 is the incident X-ray intensity, and x is the distance traveled through the matter. This equation holds only true for the case of a homogeneous specimen with a constant attenuation coefficient. As this is not the case for practical CT applications (such

as material mixes), the equation can be rewritten to take a varying attenuation coefficient into account [Carmignato et al., 2018].

$$I(l) = I_0 e^{-\int_0^l \mu(x)dx} \tag{2.4}$$

For an in-depth physical explanation of the beam-matter-interaction, which results in the attenuation, the reader is referred to the literature [Buzug, 2008; Cierniak, 2011; Krieger, 2017; Stock, 2018].

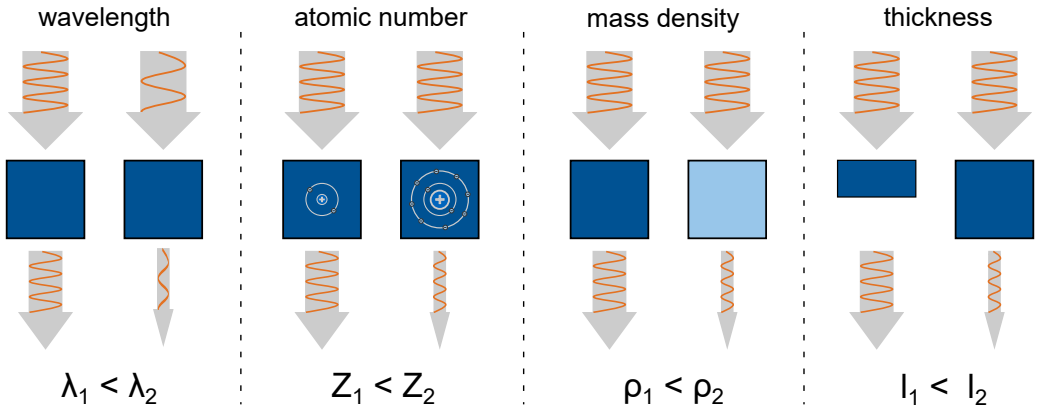


Figure 2.12 Illustration of different specimen properties and their influence on the attenuation. The attenuation increases for higher wavelengths of the radiation, higher atomic numbers of the specimen material, higher density of the specimen material or higher thickness of the specimen (adapted from [Buzug, 2008]).

In conventional CT machines, the attenuation of the X-rays is used to infer information about the specimen. For this, the radiation intensity behind the specimen is measured by a detector. Commonly, solid-state scintillators are used to convert the X-ray beam into an electrical signal. They consist of two main parts. Firstly, the scintillator medium, which captures the X-ray radiation and converts it into visible light. Secondly, the photodetector, which is placed directly behind the scintillator. The photodetector converts the visible light into an electrical signal. By arranging this setup in an array similar to the photo sensor in a conventional camera, a 2D image (projection) of the radiation intensity behind the specimen is captured [Carmignato et al., 2018].

The captured images are processed to create the 3D CT volume. This process is called reconstruction. There exist different algorithms for the volumetric reconstruction from 2D projections. The most common case in industrial CT is the reconstruction from many projections taken from predefined angles around the specimen. For this, the specimen is rotated between the X-ray source and the detector. Figure 2.13 illustrates the basic idea. It shows a specimen with its corresponding attenuation profile. Due to imperfections in the measuring setup, the edges of the specimen might not be depicted sharply but show a gradual attenuation change (see the projection on the detector in Figure 2.13). By combining multiple projections in combination with the corresponding scan angle, an increasingly finer reconstruction can be achieved. Commonly the Feldkamp algorithm is used to achieve this reconstruction [Alkadhi et al., 2011]. It uses filtered back-projection to approximate the 3D density function from a set of 2D projections [Feldkamp et al., 1984]. For a detailed mathematical description, the reader is referred to the literature [Feldkamp et al., 1984; Buzug, 2008; Podgoršak, 2010; Carmignato et al., 2018].

2.2.3 Artifacts and Scan Quality

Due to imperfections in the scanning setup and the data reconstruction, different scan artifacts can occur, which limits the scan quality. Artifacts are one of the main reasons for the challenging automation of CT analysis as they introduce gray value changes and noise, which hinder or complicate the analysis by conventional image processing techniques. In the following, the most relevant artifacts for this study are introduced. For an extensive list of artifacts and a more detailed description, see [Buzug, 2008; Schörner, 2012; Kratz, 2015; Carmignato et al., 2018; Kiefel, 2017].

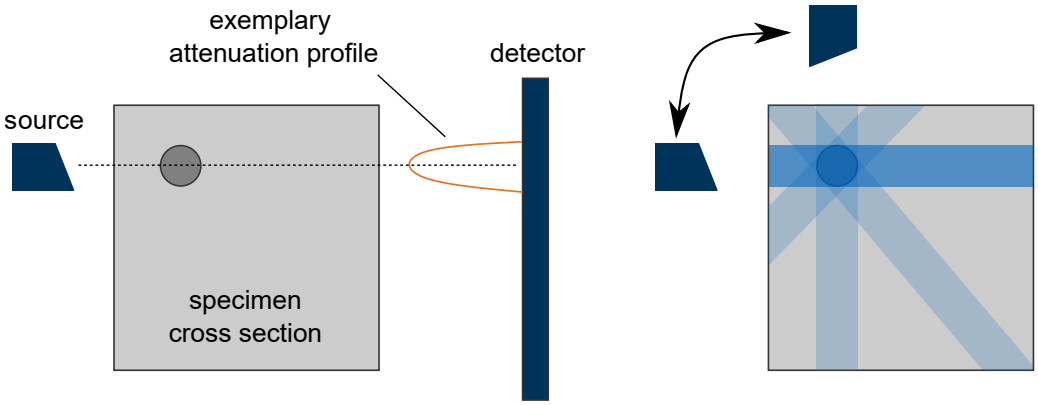


Figure 2.13 CT reconstruction of a simple quadratic specimen with a circular anomaly. The specimen is rotated relatively to the X-ray source and the detector. In the middle, an exemplary attenuation profile is shown.

Beam Hardening

Beam hardening artifacts result from the X-rays' broad energy spectrum. As discussed in Section 2.2.1, the radiation emitted by the X-ray tube has different energy levels due to the Bremsstrahlung. Photons with different energy levels interact differently with the specimen. In particular, the attenuation coefficient is frequency-dependent. Equation (2.4) does not consider this and therefore introduces an imperfection [Buzug, 2008]. Low-energy X-rays are attenuated more strongly by the specimen, while high-energy X-rays pass through more easily. Consequently, the energy spectrum of the radiation becomes richer at high-energy levels, influencing the mean intensity behind the specimen. The more low-energy radiation is filtered, the higher the mean radiation intensity. Therefore, the mean intensity measured by the detector is shifted. This non-linear relation leads to underestimating the attenuation and lowering the opacity value (see Figure 2.14). Beam hardening artifacts might be reduced by filtering the low-energy X-ray radiation with a physical filter [Kratz, 2015; Carmignato et al., 2018].

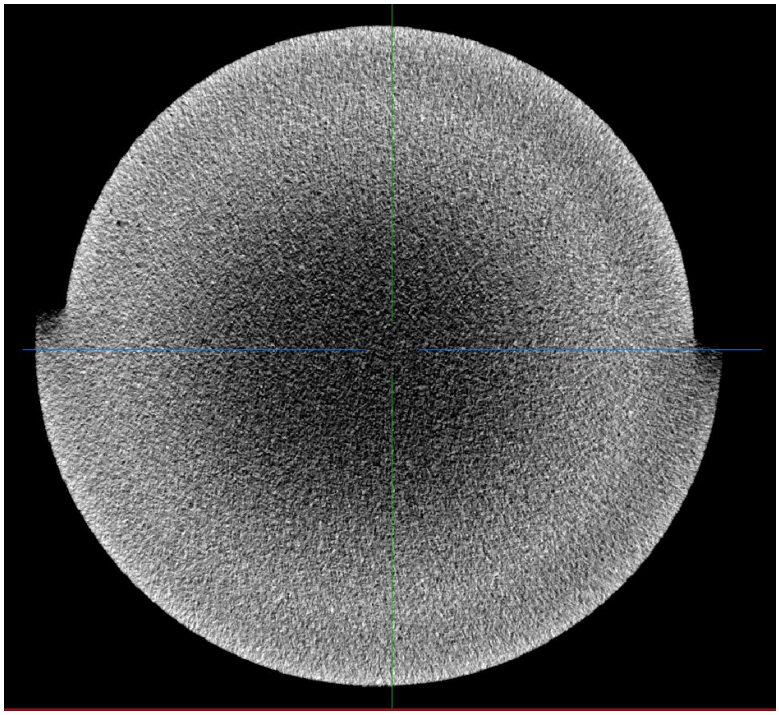


Figure 2.14 Exemplary cross-section of a specimen showing a gradient in gray values from the middle to the border due to beam hardening.

Partial Volume Effect

The partial volume effect might appear at borders in the specimen with a sudden change in density. Due to a limited number of pixels on the detector, the border might not be captured sharply. When the border does not coincide with the borders of the pixel grid on the detector, one pixel might capture information about the material on both sides of the border. This might lead to a diffuse change in gray value instead of a sharp edge (see Figure 2.15). This effect is influenced significantly by the chosen scan resolution [Buzug, 2008].

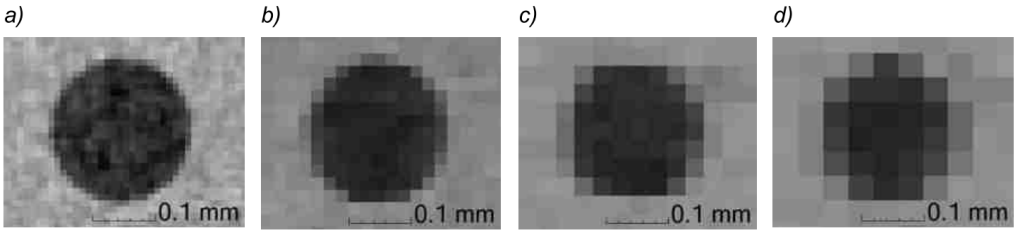


Figure 2.15 Illustration of the partial volume effect based on CT scans with varying resolution: a) 10 µm, b) 20 µm, c) 30 µm, d) 40 µm [Kiefel, 2017].

Scatter Artifacts

Scatter artifacts are caused by the scattering of the X-ray radiation by the specimen. Scattering depends mostly on the material of the specimen, its geometry, and the distance between the specimen and the detector. It exhibits similar artifacts in the gray value distribution as beam hardening, leading to a lower contrast and streak artifacts within the reconstructed volume (see Figure 2.16). It can be reduced by increasing the distance between the specimen and detector or by using special reconstruction methods [Schörner, 2012; Kratz, 2015].

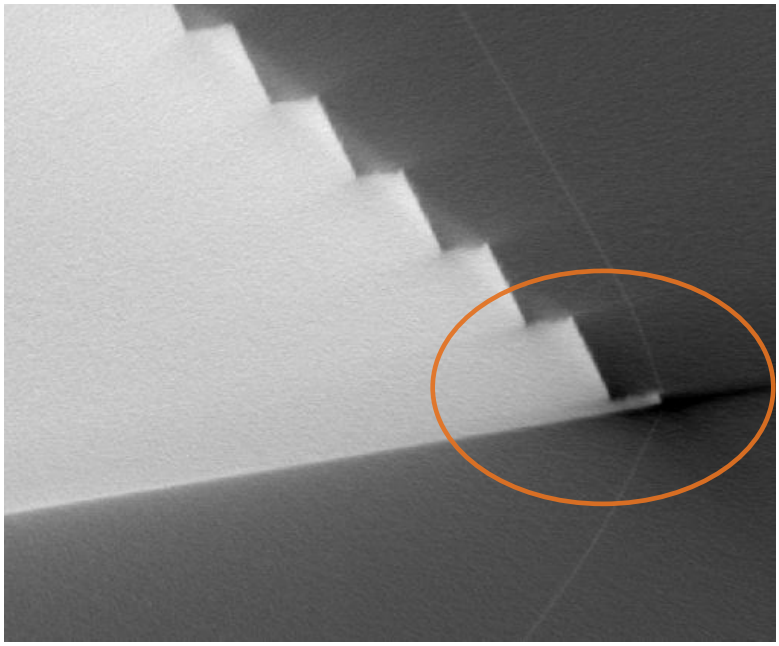


Figure 2.16 Exemplary scatter artifact visible as a darker streak at the lower right corner of a specimen.

Ring Artifact

Ring artifacts are concentric circular structures that appear as artifacts in the reconstructed volume (see Figure 2.17). They show a contrasting gray value to their surroundings depending on the cause of the specific artifact. Ring artifacts are often caused by an insufficient calibration of the detector, defective pixels, or dust on the detector. Another reason can be beam hardening. They can be avoided or reduced by regular calibrating and cleaning the detector. Another possibility is to place the specimen outside of the affected detector area [Carmignato et al., 2018].

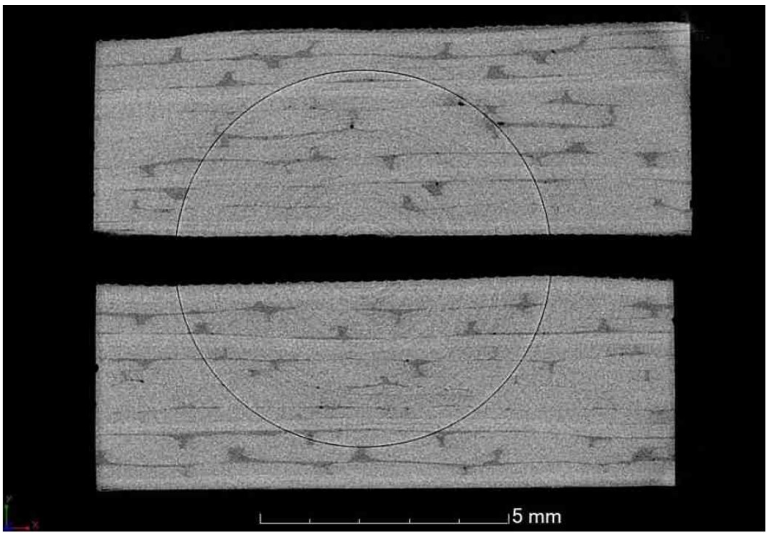


Figure 2.17 CT scan with clearly visible ring artifact in the centre [Kiefel, 2017].

2.3 Neural Networks

In this chapter, firstly, the mathematical concept underlying neural networks and their training are presented. Secondly, the basic network structure of convolutional neural networks is introduced, and their application to image analysis tasks is explained. Thirdly, the training process and associated (hyper-) parameters are presented, and the foundations for evaluating and verifying self-implemented neural networks are provided. Due to the large and fast-evolving field of deep learning research, only concepts and architectures specific to this project are discussed. For a more in-depth discourse, the reader is referred to the literature [LeCun et al., 1998; Goodfellow et al., 2016; Buduma and Locascio, 2017; Chollet, 2018].

2.3.1 Basic Model of a Neuron

The basic concept of an artificial neural network is adopted from the human brain. The human nervous system consists of approximately 86 billion neurons, which are connected by approximately 10^{14} - 10^{15} synapses [Stanford University, 16.05.2022b]. Each neuron receives an input and produces an output signal, which is passed on to other neurons. By strengthening or weakening the connection between two neurons, the influence of the signal on the next neuron can be adapted. Therefore, the signal strength and the connection to the other neurons in the human brain are developed over time by the information input. For humans, this process is called learning. For artificial neural networks, it is called training and is the basic motivation for machine learning [Gurney, 2003; Stanford University, 16.05.2022b; Goodfellow et al., 2016; Buduma and Locascio, 2017]. The analogy between artificial neural networks and the human nervous system is coarse but holds for the basic motivation of training/ learning. For a more detailed explanation of the actual human nervous system, see [London and Häusser, 2005; Brunel et al., 2014].

Figure 2.18 shows the mathematical model of an artificial neuron. The artificial neuron takes the weighted sum of i inputs I_i and adds a constant bias b. The weight of each input is referred to as w_i. The sum is passed into the activation function $f(z)$ which returns the output Y.

$$Y = f(\sum(I_i w_i) + b) \tag{2.5}$$

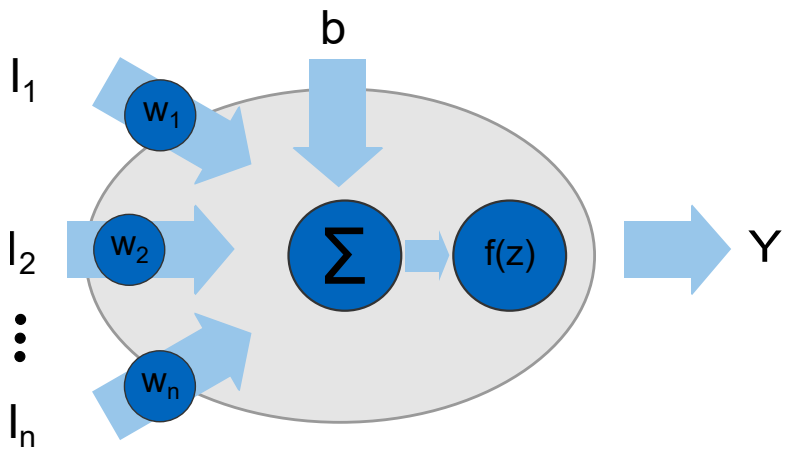


Figure 2.18 Illustration of the underlying mathematical model of an artificial neuron. With the inputs I_i, the weights w_i, the bias b and the output Y.

Weights

The weight of an input determines its influence on the overall input sum. Hence, it defines the influence of a preceding neuron (or input data point) on the current neuron. The weights of a neuron are altered during the training to adapt to the given task. The principle by which the weight is adopted is explained in Section 2.3.2. The initial values of the weights can be preset when the neural network has been initialized and trained before. In this case, the weights are referred to as pre-trained, which are often used in transfer learning tasks. This is desirable in most cases as it improves the training duration and performance. When no prior knowledge about the weights is available, the weights should be initialized randomly with small numbers. This facilitates the training compared to the initialization with a constant [Stanford University, 16.05.2022c].

Bias

Similar to the weights, the bias is also altered during the training process. The bias is a constant added to the inputs' weighted sum. It is, therefore, used to shift the activation function by a constant value. For example, when the input to the activation function is supposed to be one, when all inputs to the neuron are zero, the bias is set to one. The bias is usually initialized to zero when no prior knowledge is available [Stanford University, 16.05.2022c].

Activation Function

The activation function $f(z)$ introduces non-linearity to the neuron. So far, the neuron acts as a simple algebraic adder that takes the sum of all weighted inputs and the bias. It can be shown that a network consisting of only linear neurons can be expressed as a network with no hidden layers. As shown later, hidden layers are crucial to learning complex features. Therefore, the ability of neurons to represent non-linear behavior is essential for many learning tasks [Buduma and Locascio, 2017]. Three major activation functions are used in practice to introduce non-linearity in their computation. They all take the sum of the weighted inputs and the bias and return the activation value, which is the output of the neuron.

The sigmoid function is defined as [Buduma and Locascio, 2017]:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.6}$$

The function's behavior is shown in Figure 2.19. For large negative values, a very small value is returned. Around zero, an s-shaped shift is observed, and for large positive values, the output is close to 1. By introducing additional constants to the function, the shape and location of the S-shift can be altered [Gurney, 2003]. Commonly, the basic sigmoid function is used.

The tanh function is similar to the sigmoid function but returns a zero-centered output between -1 and 1 (see Figure 2.19). Therefore, it is often preferred over the sigmoid function. It is defined as [Buduma and Locascio, 2017]:

$$f(z) = tanh(z) \tag{2.7}$$

The most commonly used activation function in computer vision is the ReLU function (rectified linear unit). It returns zero for all negative values and forwards the input for all positive input values (see Figure 2.19). It is defined as:

$$f(z) = max(0, z) \tag{2.8}$$

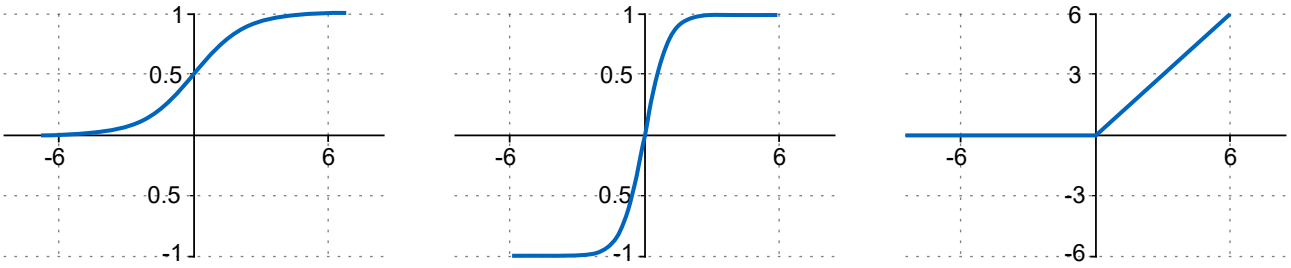


Figure 2.19 Three common activation functions applied in neural networks. From left to right: sigmoid function, tanh function, ReLU function.

2.3.2 Neural Network Training

The basic neuron introduced in Section 2.3.1 can be used to perform elementary operations. For the neuron to return the desired value, the weights and biases have to be adopted. This is done in a training process. During training, the neuron is shown a predefined set of data points together with the desired result. The targeted result is called ground truth and represents the desired output of the neuron. Hence, the goal of the training process is to minimize the error E (e.g. the mean square error) between the actual neuron output $f(z)$ and the ground truth t. With i denoting the number of training data points, it can for example be written as [Buduma and Locascio, 2017]:

$$E = \frac{1}{2} \sum_{i} (t_i - f_i(z))^2 \tag{2.9}$$

To achieve this, in the first step, the data is passed into the neuron, and the resulting output is compared to the ground truth data. The closer the output is to the ground truth, the smaller the mean squared error and the better the model. As $f(z)$ depends on the weights w_i and the bias b, the error function E also depends on these parameters. Therefore, the goal is to find a set of parameters w_i' and b' which minimizes the error function. As the activation function is non-linear, this equation cannot be solved analytically [Buduma and Locascio, 2017]. Instead, numerical optimization is used to minimize the overall error by gradient descent. The choice of error function is use-case-specific.

Gradient Descent

For simplicity reasons, the general concept of gradient descent is explained on a single neuron with only two weights and no bias. Hence, the error function only depends on w_1 and w_2. Figure 2.20 shows an exemplary error surface for such a setup. The optimal set of weights is reached when the error is the lowest, hence when the lowest point on the slope is reached. Starting from a random point (as the weights are initialized randomly) on the graph, this is achieved by following the steepest descent at this point. Mathematically, this can be achieved by evaluating the gradient vector at this point and calculating the next point on the graph based on it. By repeatedly following the steepest descent, a minimum is reached.

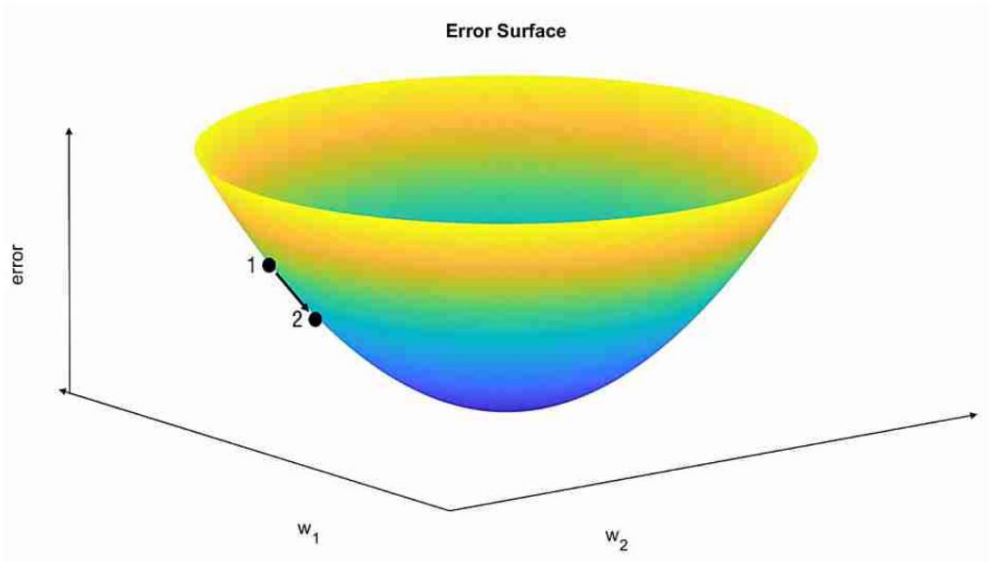
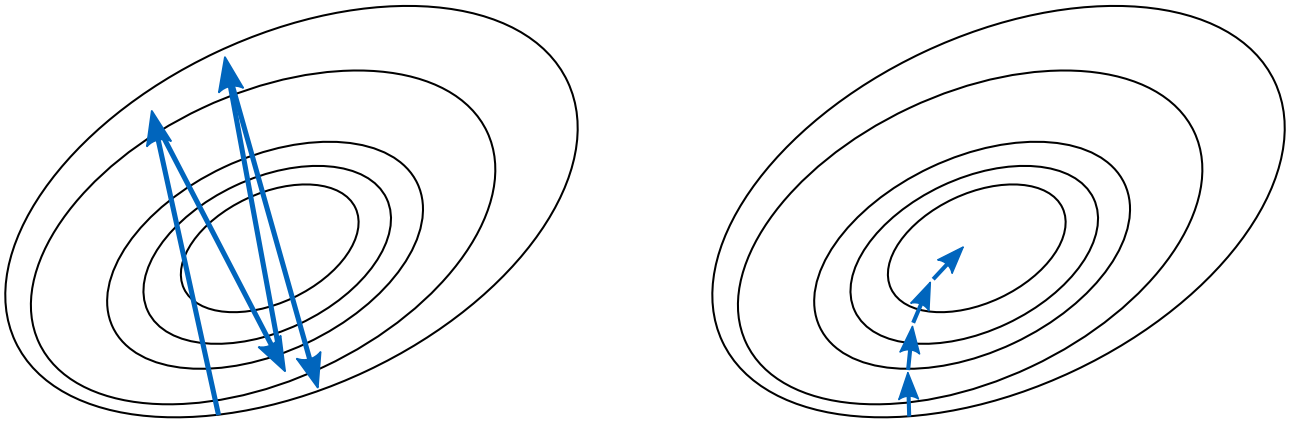


Figure 2.20 Illustration of the error surface for a single neuron with only two weights w_1 and w_2 and the gradient descent method.

For more complex error surfaces, an improved gradient descent might be necessary. The step size for each iteration can be adapted by multiplying a factor by the gradient. This factor is called the learning rate and represents an important hyperparameter of the training process. The learning rate influences the training speed and performance. A high learning rate might speed up the training as it allows for a fast marching adaption of the weights, but at the same time, it might hinder the algorithm's convergence to a minimum. On the other hand, a low learning rate might slow down the training process considerably. This is illustrated by Figure 2.21. Therefore, the choice of the learning rate is non-trivial and might have to be adapted per training or even during the training process [Buduma and Locascio, 2017].

Feedforward Neural Network

The feedforward neural network (FNN) is introduced to extend the idea of gradient descent to a more complex network. The idea of an artificial neural network is the successive arrangement of neurons. This allows for the representation of more complex behaviors. Commonly, the neurons are arranged in layers as shown in Figure 2.22. In a FNN, all neurons are connected in one direction. In particular, no feedback loop exists to prior neurons or within one layer. The input layer takes the data input and returns an output, which is passed on to the next layer. The successive layers are called hidden layers, as they are in between the input and output layers. The number of hidden layers is network-specific. The output layer receives the output of the last hidden layer and returns the output of the model. It can be mathematically proven that an FNN with one hidden layer and sigmoid activation functions can approximate any continuous multivariate function to any accuracy [Cybenko, 1989; Du and Swamy, 2019].



(a) A high learning rate might hinder the convergence of the gradient descent algorithm.

(b) A low learning rate might slow down the training process considerably.

Figure 2.21 Influence of the learning rate on the training process (adapted from [Buduma and Locascio, 2017]).

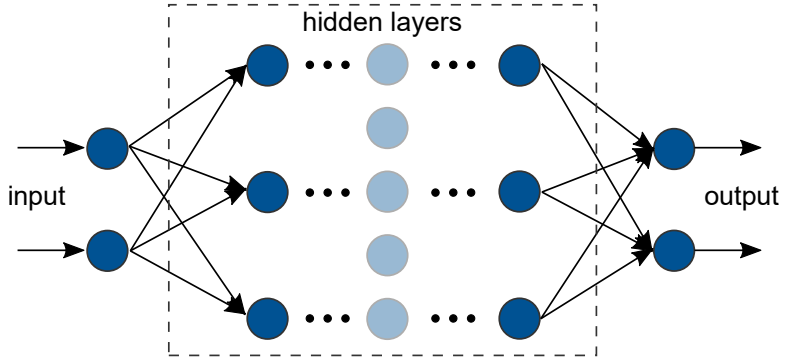


Figure 2.22 Illustration of a feedforward neural network with two input neurons, an arbitrary number of hidden layers, and two output neurons.

Backpropagation

The large number of weights and biases in an FNN demands an extension of the gradient descent approach to a multidimensional space. The goal is to minimize the model's overall error, considering all weights and biases. As each neuron affects multiple following neurons, the influence of each neuron must be considered. Backpropagation aims to solve this problem by iteratively calculating the loss derivative per layer, propagating the loss received at the output backward through the network [Rumelhart et al., 1986]. The general concept follows the explanation for the gradient descent. For a more in-depth investigation and mathematical extension into the multidimensional space, the reader is referred to [Rumelhart et al., 1986; LeCun et al., 1998; Buduma and Locascio, 2017; Du and Swamy, 2019].

Loss Functions

In order to minimize the error of the network by backpropagation, a loss function has to be defined. The definition of the loss function is task-specific as the network is pushed to optimize its performance with respect to this loss. There exists a variety of different loss functions. The loss functions relevant to this study are introduced in detail in Section 4.4. In the following the generic concept of the training and validation loss are introduced.

While the training loss function is used to train the model, it is not sufficient for an adequate assessment of the model's performance. In particular, the network might perform very well on the training dataset but might not be able to generalize this behavior to other data. Figure 2.23 visualizes this phenomena. The goal of the network is to optimally predict a curve based on the given points in green. The points can be

regarded as the given data points. It is not possible to obtain and train the network with all existing data points. Therefore, the objective is to find a curve that can be generalized to further data points (shown in orange). Two curves are fitted to the plot exemplary. The polynomial curve fits perfectly to the training data points and, therefore, results in a training loss close to zero. On the other hand, the linear function shows a non-perfect fit as it only approximates the given data points to a certain degree. Hence, the training loss is higher. Evaluating the network performance solely based on the training loss, the polynomial fit would be regarded as the superior model. However, considering the additional data points, the linear fit would represent the actual data better [Buduma and Locascio, 2017].

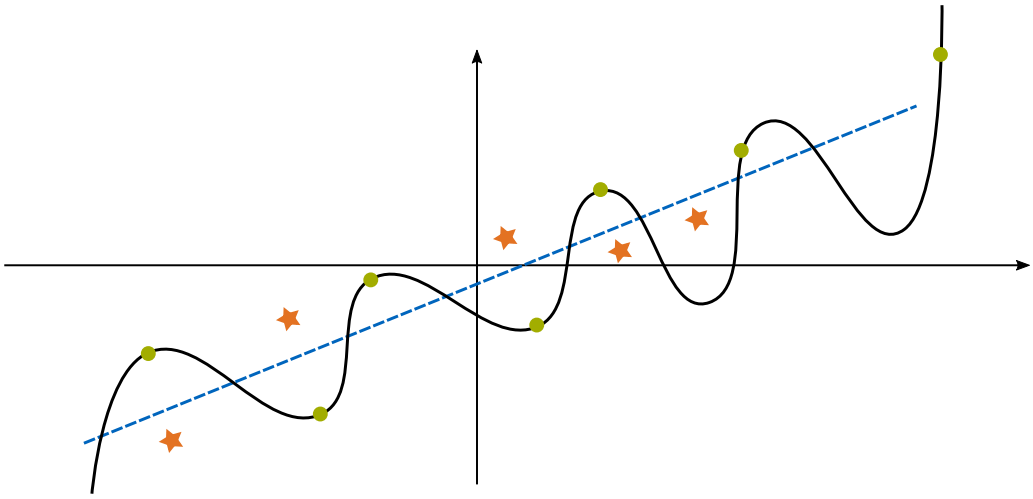


Figure 2.23 Example of two different model outputs with one model overfitting to the training data (solid black curve) and one model generalizing well by simple linear fit (dotted blue line). The green dots represent training data points. The orange stars show additional data points not known during training (adapted from [Buduma and Locascio, 2017]).

This illustrates two requirements for the training. Firstly, it shows the importance of the data quality and, particularly, that the training data has to represent a relevant data variety for the given use case. Secondly, even for a good training dataset, acquiring every possible data point is not feasible, and it is hence only based on a data subset of reality. Therefore, it is important to interpret the ability of the network to generalize. An additional metric, the validation loss, is introduced to achieve this. The validation loss is calculated similarly to the training loss but solely on additional data not used for training (the additional data points in Figure 2.23). An exemplary training and validation loss curve is shown in Figure 2.24. While the training loss decreases over time, the validation loss decreases only up to a certain number of training steps and then rises again. A common explanation for this behavior is the overadaption of the model to the training data, leading to an inferior generalization. This phenomenon is called overfitting. Possible countermeasures include an increased data variety either by an increased data gathering or an artificial data augmentation [Buduma and Locascio, 2017]. Common data augmentation techniques for images include geometric augmentation, intensity augmentation, and the addition of noise. A more detailed list and analysis can be found in literature [Shorten and Khoshgoftaar, 2019].

2.3.3 Convolutional Neural Network

Convolutional Neural Networks (CNN) are a particular class of neural networks that use convolutions. Over the last decade, CNNs have shown great potential in analyzing 2D and 3D images. They were first introduced by LeCun et al. [1989] to interpret human handwriting automatically.

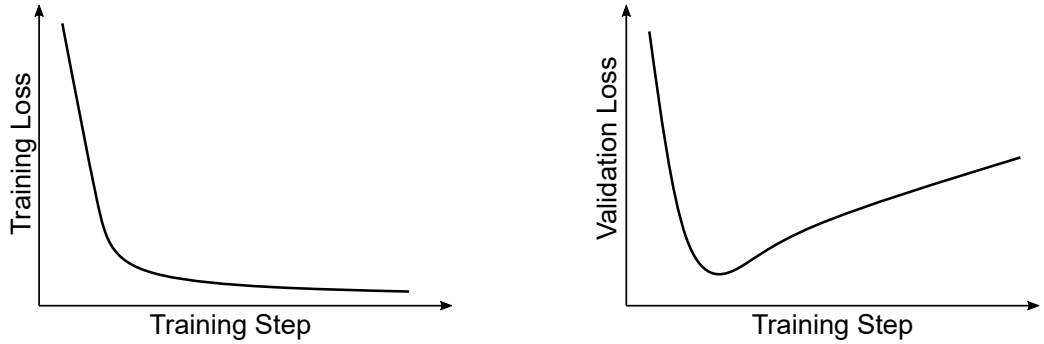


Figure 2.24 Example of the training and validation loss curve for an overfitting model.

Convolutional Layer

The basic concept of CNNs is the extraction of local features. To achieve this, CNNs use so-called filters. A filter is sensitive to a specific feature and returns an output depending on this feature. Figure 2.25 shows an example of a possible filter setup. For simplicity, the binary input image (on the top) contains a simple form of vertical and horizontal lines. Two possible filters are shown in the middle. One filter is sensitive to vertical lines (left), while the second is sensitive to horizontal lines. The filters are slid across the input image and return a positive output when the filter feature fits the input structure, e.g., when a vertical/ horizontal line is detected. The output of the filter is called a feature map. In the example in Figure 2.25, the two feature maps for the corresponding filters are shown at the bottom.

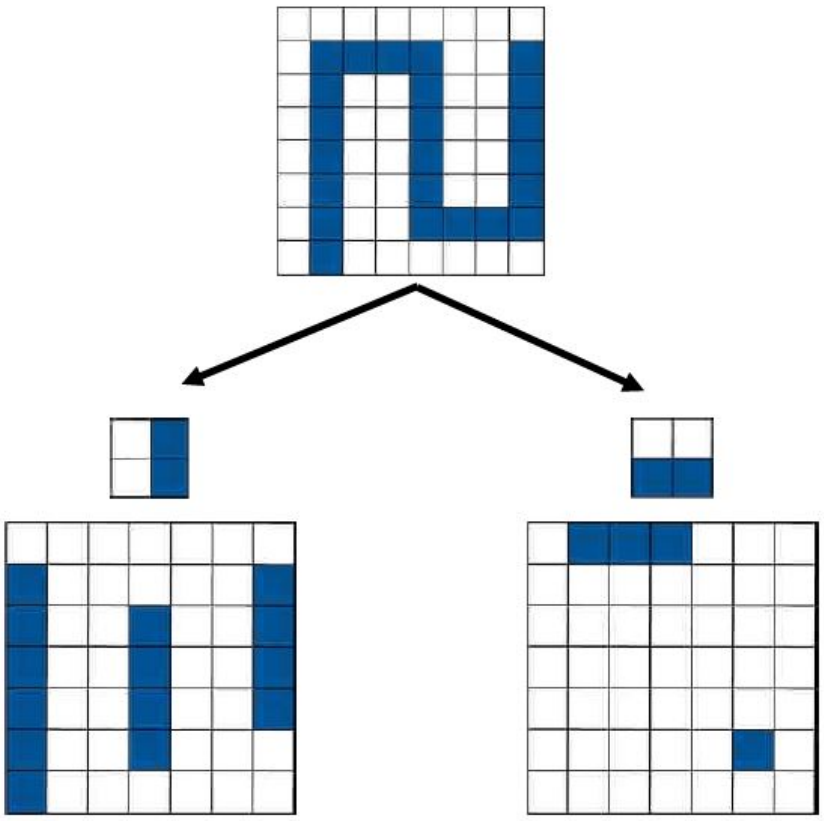


Figure 2.25 Illustration of two filters: one reacting to horizontal lines, the other to vertical lines. The output is a feature map shown below (adapted from [Buduma and Locascio, 2017]).

As for other feedforward neural networks, the output of one filter layer can be passed on to a subsequent layer. In the given example, the next layer would, therefore, receive the information about the existence of

vertical and horizontal lines. Hence, adding additional layers can represent more complex features from the input image, e.g., points at which horizontal and vertical lines touch. Transferring this example to more complex tasks, e.g., recognizing the TUM logo, a possible set of features is shown in Figure 2.26.

In practice, the filter sensitivity to a certain feature is learned during training by adapting the weights and biases of the corresponding filter. In a convolutional layer, this is achieved by combining multiple neurons into a cluster that is connected to specific neurons in the previous layer. Additionally, certain clusters within one layer share the same weights, allowing the detection of the same features. This has the positive effect of drastically reducing the number of trainable parameters, which in turn facilitates training. At the same time, this leads to a translation invariance of the model as a feature is detected independently of its position in the image [LeCun et al., 1989; Buduma and Locascio, 2017; Chollet, 2018; Stanford University, 16.05.2022a].

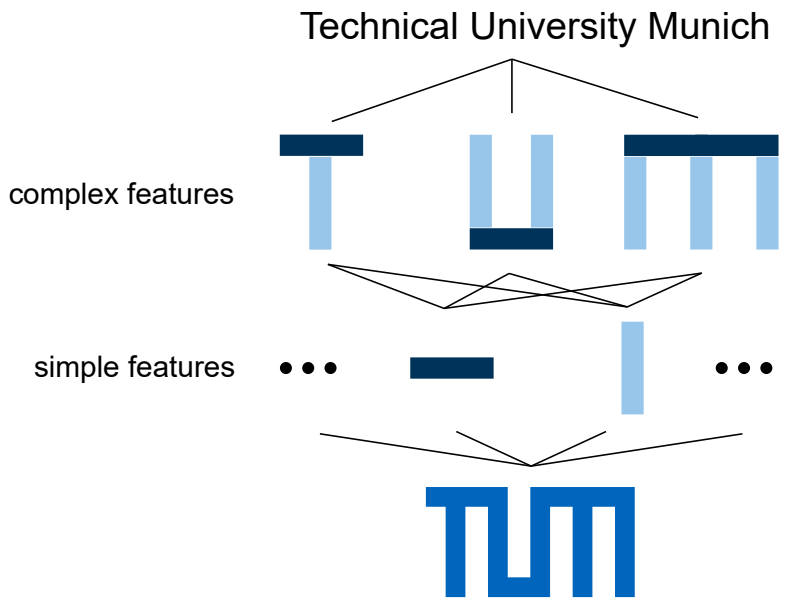


Figure 2.26 Use of multiple layers to combine simple features into more complex features to detect a complex geometry (e.g.the TUM logo) (adapted from [Chollet, 2018]).

Pooling Layer

Besides convolutional layers, common CNNs often implement Pooling Layers. The purpose of pooling layers is the progressive reduction of spatial size and, thereby, a reduction of trainable parameters and computation requirements. There exist different pooling strategies. The most common pooling method is maximum pooling (max-pooling). A max-pooling filter returns the maximum values of all its inputs. The filter size and the stride by which it is moved are important hyperparameters determining the data reduction degree. Other pooling operations are average pooling or L2-norm pooling [Stanford University, 16.05.2022a].

Upsampling Layer

In recent years, Fully Convolutional Networks have shown great potential in image segmentation tasks [Long et al., 2015; Ronneberger et al., 2015]. In contrast to prior network architectures, which used fully connected layers as final layers, fully convolutional networks use so-called up-sampling or deconvolution layers. As the name suggests, these layers act inversely to convolutional layers. They take a feature map as input and return a number of (less complex) feature maps. Therefore, the information about a complex feature is passed back to a map of the same spatial size as the input. The result is a label map of the same size as the input with a prediction for each pixel for the corresponding class [Long et al., 2015].

3 State of the Art

3.1 Convolutional Neural Networks for Meltpool Characterization

3.1.1 Development Framework

In recent years, the field of medical imaging has experienced a significant surge in research activity driven mainly by artificial intelligence. In particular, the research and development of state-of-the-art CNNs has shown great potential [Liu et al., 2019; Daugaard Jørgensen et al., 2022].

The analysis of CT and OM data shows some parallels with the challenges faced in the evaluation of medical images, e.g., computed tomography images of a patient. Firstly, the amount of gathered data calls for an automatic evaluation by algorithms. Secondly, all three data sets can be represented by 3D volumes and the desired output is often a segmentation of such 3D volumes. Hence allowing for the use of 3D CNNs. Thirdly, the available ground truth data is limited as manual labeling of the data is time- and cost-intensive [Ronneberger et al., 2015; Cicek et al., 2016]. While those parallels allow for the adoption of newly developed network architectures from medical imaging to the given task, the differences in domain call for task-specific optimizations, in particular, concerning the validation and testing of the trained models.

The open-source project MONAI aims to become the standard framework for developing state-of-the-art artificial intelligence applications in medical imaging [Nic Ma et al., 2021]. It is based on PyTorch [Paszke et al., 2019] and incorporates benchmark architectures such as U-Net [Ronneberger et al., 2015] and improvements of it [Kerfoot et al., 2019]. Additionally, it provides the possibility to adapt those models to specific use cases. Therefore, it is used as the implementation framework for the CNNs described in this thesis.

3.1.2 U-Net

As described above, the use of artificial intelligence in medical imaging is an active field of research. In the last decade, the number of peer-reviewed articles per year has been constantly growing, with over 1600 publications in 2019 alone [Zhang et al., 2021]. Introduced in 2015, the U-Net architecture outperformed the then state-of-the-art architectures for image segmentation by a significant margin [Ronneberger et al., 2015]. It uses the idea of a fully convolutional neural network [Long et al., 2015] and consists of a contracting path and an expanding path. Figure 3.1 shows the general architecture. The original paper's contracting path consists of a series of 3x3 convolutions, rectified linear units, and max pooling operations. The expanding branch uses a combination of upsampling and up-convolution units. Additionally, the cropped output feature maps of the intermediate steps in the contracting path are fed to the corresponding steps in the expanding branch. This setup allows for the retention of high-resolution features by the introduced "shortcuts" from the downsampling to the corresponding upsampling unit. At the same time, the global context information can be captured and propagated well by passing through the entire network with a high number of feature channels. As the network does not contain any fully connected layers, it allows for the segmentation of almost arbitrarily large images. For a computationally efficient implementation, the entire image can be cropped to smaller tiles, which are fed through the network and then reassembled with a pre-defined overlap [Ronneberger et al., 2015].

The basic architecture described above has been extended to 3D volumes and used as a baseline by different authors [Falk et al., 2019; Cicek et al., 2016]. Kerfoot et al. [2019] introduced residual units to the U-Net architecture. While more adaptions to the basic architectures have been proposed Isensee et al. [2021] suggest that the correct choice of details in the method configuration and pre-processing can be even more important than the exact network architecture. This hypothesis was already put forward

by Litjens et al. [2017]. After reviewing over 300 publications in the field of deep learning in medical image analysis, they concluded that the identical architecture can result in widely varying results. Expert knowledge about the data domain greatly impacts the final performance as it determines important pre- and post-processing steps such as data augmentation. Nonetheless, the general network architecture and the input size were also identified as relevant design choices. A clear impact of the hyperparameter choice could not be shown and is believed to be of secondary importance [Litjens et al., 2017].

The "nnUnet" builds on these results and aims to be a baseline for biomedical segmentation tasks by extracting a "fingerprint" of the given dataset and designing a fitting training pipeline using pre-defined heuristic rules. This approach could outperform the prior state-of-the-art network in 33 out of 53 medical segmentation tasks [Isensee et al., 2021]. U-Net and its different adaptations can therefore be seen as the de facto standard for medical segmentation tasks, as can also be seen by the "Medical Segmentation Decathlon". The Medical Segmentation Decathlon consists of ten different medical datasets with differing properties [Antonelli et al., 2021]. The top 15 methods were all based on the U-Net architecture [Isensee et al., 2021]. Therefore, this thesis will use the U-Net as the baseline architecture for segmentation.

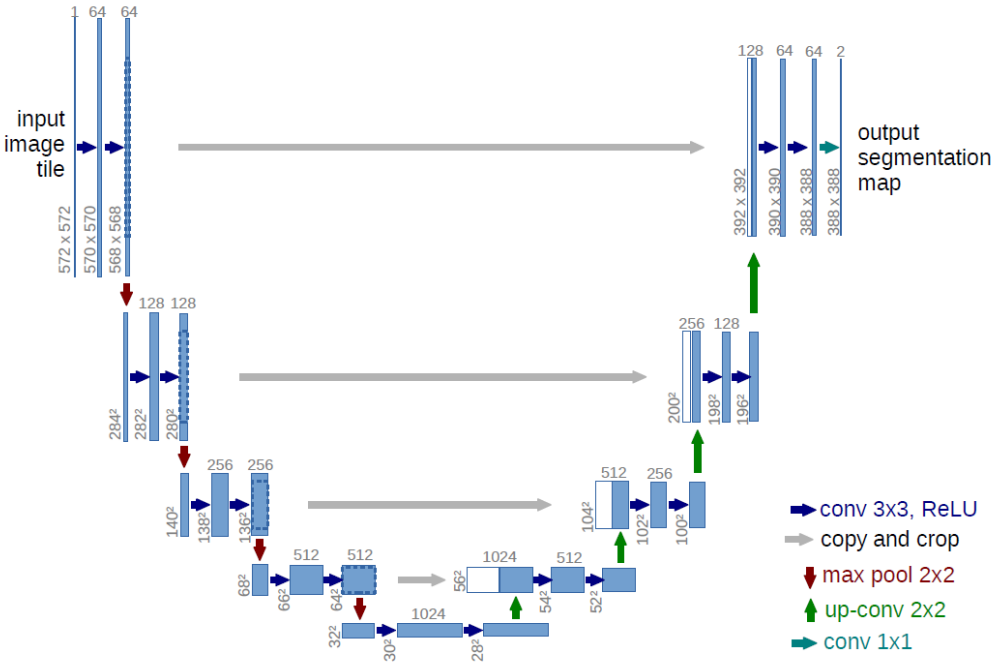


Figure 3.1 Basic U-Net architecture as introduced [Ronneberger et al., 2015].

3.2 Online Monitoring

The process monitoring of the LPBF process is an active field of research for academic research groups, industry, and agencies. There exists a wide variety of different approaches ranging from off-the-shelf sensor systems to gather basic process parameters to highly complex research systems to better understand meltpool dynamics. In the following, the focus is placed on monitoring systems that aim to predict the quality of the finished part. Multiple review papers have been published to summarize the current research findings and provide an overview of the state of the art [Everton et al., 2016; Grasso and Colosimo, 2017; Goh et al., 2020; Meng et al., 2020; Wu et al., 2021; Mahmoud et al., 2021]. Three key findings can be extracted from those.

Firstly, online monitoring has been identified as a critical component to increase the confidence in and understanding of the AM process. It is predicted that resolving this bottleneck can widely promote the industrial breakthrough of LPBF, i.e., in safety-critical applications.

Secondly, a variety of different sensors have been installed and tested to monitor the LPBF printing process. There is no clear recommendation on which sensor setup is most promising to capture relevant process data, but most setups concentrate on optical systems to monitor the printing. Data fusion of different sensors is expected to increase the monitoring performance.

Thirdly, while extensive research is conducted concerning sensor setups and different data-gathering hardware, the analysis of the obtained data and correlation with other data sources has only recently become the focus of research. In particular, using machine learning to analyze monitoring data has shown promising results [Razvi et al., 2019; Goh et al., 2020; Mahmoud et al., 2021].

3.2.1 Sensor Setups

As pointed out above, extensive research has been conducted with respect to the best sensor setup for online monitoring. The sensor systems can be categorized by the physical property they monitor. The most common systems use optical, temperature, or acoustic sensors. Furthermore, eddy current systems, X-ray-based systems, and even more complex research setups have been tested [Everton et al., 2016; Grasso and Colosimo, 2017; McCann et al., 2021].

Acoustic Techniques

Most acoustic systems use the sonic or ultrasonic waves generated by the printing process itself. Possible origins are the laser melting of the powder, the recoater movement, or the cracking of the printed part. The generated waves propagate through the manufactured part and can be recorded by air or by coupling the sensor to the buildplate. Systems can be categorized into ultrasonic (UT) and acoustic emission systems, with both sensor principles being adopted from classical NDT applications [McCann et al., 2021].

UT is widely used for non-destructive testing of metal parts, i.e., in the aerospace industry. It uses ultrasound waves to gather information about the internal properties of the specimen and can be used to detect voids or porosities [Schiebold, 2015a]. Initial research suggests that UT as an online monitoring system might be able to detect porosity in the specimen, but further development is required, i.e., concerning the monitoring of more complex geometries and the stabilization of the inspection setup (e.g., robust coupling of the sensors) [Dillhoefer et al., 2014; Rieder et al., 2016].

Acoustic emission sensors capture a broader frequency range than UT sensors. By analyzing the acoustic spectrum, the global process quality can be estimated, e.g., by applying an artificial neural network, the quality of a specimen could be classified with approximately 80% accuracy as poor, medium, or high based on altered process conditions [Shevchik et al., 2018; Eschner et al., 2020]. Additionally, acoustic emission can be used to monitor the crack initiation or growth within a specimen [McCann et al., 2021]. While being able to detect specific local defects (e.g., cracks) and specific global part qualities, the applicability of acoustic emission for broader monitoring tasks remains a research topic. In particular, the transfer to more complex geometries and the robust analysis of the sensor data hold potential for further improvement [Eschner et al., 2020].

Optical Tomography

The term optical tomography has multiple meanings in different disciplines. While it describes the method to probe highly scattering media in medicine, it also describes the use of light to image fast phenomena, e.g., gas combustion [Arridge, 1999]. In the field of LPBF, Optical Tomography (OT) primarily refers to a commercially available sensor system by EOS GmbH. An in-depth analysis and development of the system can be found in Ladewig [2019]. It was industrially introduced by MTU Aerospace and Carl Messtechnik and uses a camera with a near-infrared filter [Bamberg et al., 2016; Carl, 2015]. In the published setup, the camera captures images of the entire build platform with a spatial resolution of $125\,\mu m$ per pixel with a frame rate of 10 frames per second. The acquired images are superimposed to create one image per layer. This image can be regarded as a process map correlating to the light (in the near-infrared wavelength range) emitted by the process per layer [Fuchs and Eischer, 2018; Listl and Orye, 2019]. As

the near-infrared bandpass filter largely filters the monochromatic laser light, the recorded light can be attributed mainly to the thermal radiation caused by the meltpool [Goegelein et al., 2018]. The general setup is shown in Figure 3.2.

The system is used to monitor process deviations, e.g., it can detect deviations in the laser power, scan speed, and hatch distance greater than 5 %. It could be shown that a strong deviation of these parameters leads to a degradation in part quality, i.e., to an increase in porosity [Fuchs and Eischer, 2018]. A correlation of individual defects in the specimen (detected by CT) with indications in the OT showed that one specific type of defect (lack of fusion) with a high probability leads to an indication in OT. On the other hand, not every indication in the OT was found to correlate to an actual defect. Hence, the system is able to detect the specific type of lack of fusion with a high probability at the cost of misclassifying defect-free parts as anomalous [Goegelein et al., 2018]. Furthermore, preliminary studies suggest that the OT can detect certain foreign particles such as specific inclusions [Ladewig, 2019].

The OT system was found to capture relevant process characteristics. To enhance the system's capabilities, i.e., the interpretation of the gathered data, a correlation with other data sources (e.g., post-process NDT such as CT) and a Big Data approach is proposed [Ladewig, 2019].

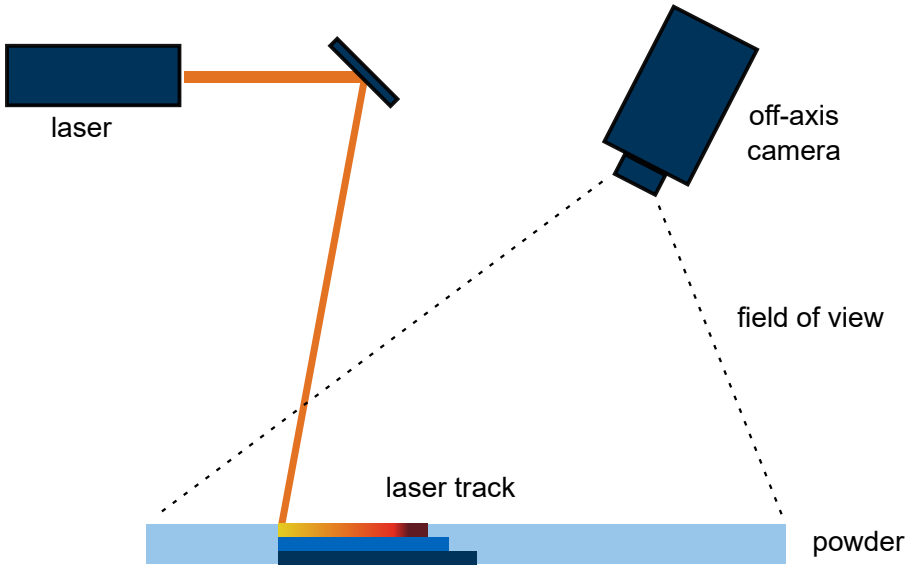


Figure 3.2 Schematic setup of an off-axis system such as the Optical Tomography. The camera captures the entire buildplate during the printing process and does not move with the laser beam (adapted from [Zenzinger et al., 2015]).

Meltpool Monitoring

The meltpool dynamic has been identified as one of the most important factors for the process and part quality. Hence, meltpool monitoring has been the subject of extensive research in recent years. It represents the highest level of detail of the presented online monitoring methods, which imposes special requirements on the sensor setup (i.e., the ability to monitor small regions at a high sampling rate due to a high dynamic) [Grasso and Colosimo, 2017]. In broader terms, the systems mentioned above, such as the OT, also monitor the meltpool as an important process parameter. Therefore, in the following, the term "meltpool monitoring" is restricted to on-axis optical or thermographic systems. Such systems are installed on-axis with the laser using the same scanner setup as the laser of the printer. Figure 3.3 shows an exemplary setup.

The basic process phenomena and meltpool mechanisms were described in Section 2.1. Although the deterministic prediction of defects based on the meltpool characteristics has not been shown yet, research suggests a strong physical relationship between the meltpool characteristic and the quality of the printed part [Lane et al., 2020]. In lack of a deterministic model the data driven model presented in this work should provide valuable empirical insights into this relationship.

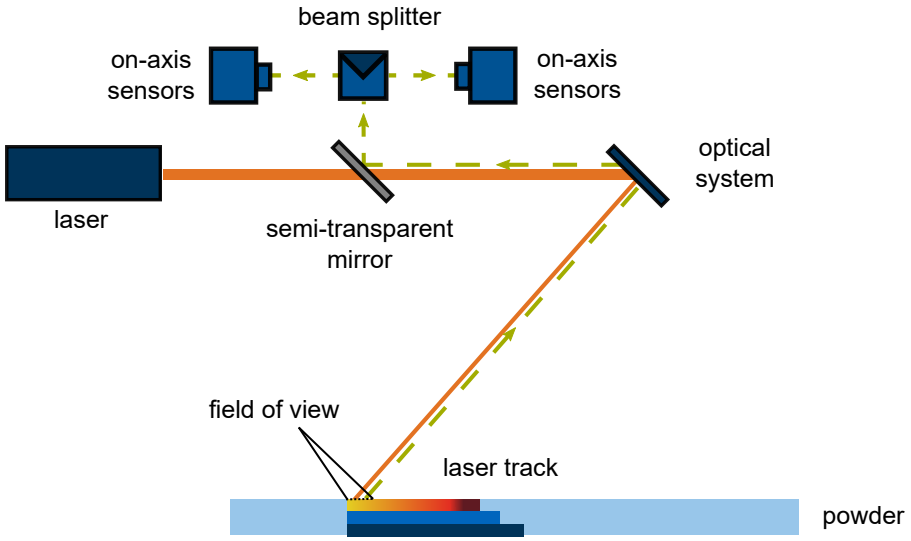


Figure 3.3 Schematic setup of an on-axis system for meltpool monitoring. The sensor is installed in parallel to the laser and captures the radiation of the meltpool, which is redirected through the laser scanner and a semi-transparent mirror (adapted from [Betts and Jacquemetton, 2019]).

The general idea in monitoring the meltpool radiation (either in the visible or thermal spectrum) is the relationship between energy input in the LPBF process and the radiation emitted by the meltpool. The total energy input by the laser can be assumed to be absorbed into the material or reflected off the surface. The absorbed energy can be divided into several factors, such as thermal conduction into the substrate, the vaporization of powder, and radiative and convective heat loss [Lane et al., 2020; Simonds et al., 2018]. The relative influence of these factors depends on the particular material and process conditions and will result in variations of the meltpool [Bidare et al., 2018]. These variations include the meltpool size, the radiation intensity of the meltpool, and the wavelength of the radiation.

By installing a semi-reflective mirror in the optical path of the laser, Craeghs from the KU Leuven showed the general working principle of the on-axis monitoring systems. More specifically, mapping the measured data to the spatial location of interest showed great potential [Craeghs et al., 2010, 2012]. A prominent commercial system for meltpool monitoring is the *Sigma Additive Solutions* (SGLB) *PrintRite3D* system. The sensors by *Sigma Additive Solutions* are used as raw data collectors in this work. In contrast, the data processing and analysis are developed independently from *Sigma Additive Solutions*. In Betts and Cola [2018a], the authors suggest a correlation between input energy, the mechanical properties of the sample, and the meltpool emission measured by a broadband photodiode. The study evaluated the photodiode signal and mechanical properties globally per sample without considering local effects. A second study performed a parameter study concluding that the *Sigma Additive Solutions* system can support the monitoring of desirable process parameters. In particular, the process window defined by laser power and speed and their influence on the part density might be observable by the *PrintRite3D* system [Betts and Cola, 2018b; Megahed et al., 2019]. In a further study, the authors compared meltpool measurements with CT scans qualitatively [Betts, 2019]. While some rough comparisons were performed qualitatively, the authors do not provide actual analysis results or quantitative interpretations. Further studies investigate the calibration of the photodiodes by a tungsten strip lamp [Diehl et al., 2022] and the use of machine learning (decision tree) to analyze the monitoring data [Frye et al., 2020].

While the general working principle of on-axis meltpool monitoring has been shown by different authors, its application to detecting anomalies in the finished part could not be demonstrated convincingly. Nevertheless, the relevance of the measured process signatures (i.e., the meltpool characteristic) suggests that meltpool monitoring is a promising data acquisition method. Additionally, different industrialized sensor setups already exist for reliable data gathering. Hence, it was chosen as a good data acquisition baseline for the presented research.

Other Research Setups

Besides the most commonly used sensor setups, there exists a variety of research setups that capture different physical properties of the printing process.

Bartlett et al. [2018] installed a long wave infrared camera overhead of the build chamber. The setup is similar to the OT system described above but uses a camera that monitors the temperature range of $-20\,°C$ to $650\,°C$ with a low frame rate of $7\,Hz$. The images were evaluated using conventional image analysis tools. The existence, location, and size of actual defects were determined by post-process cross-sections. The system was able to detect 82% of lack of fusion defects, demonstrating that infrared thermal signatures correlate well with actual defects. Moreover, the results strongly indicate that subsurface lack of fusion defects can be detected. The defect size strongly influences the detectability in IR images. While defects around $500\,µm$ were detected to 100%, the detection rate drops to 50% for defects around $50\,µm$. This might be improved by an increased resolution (current resolution $195\,µm$x$260\,µm$) of the IR camera. The results indicate that the thermal signature of a defect is larger than the defect itself, as the current system can detect defects smaller than its resolution. The system was not able to detect keyholes. The authors suspect that the formation of such pores more than $150\,µm$ below the surface poses a challenge for surface infrared inspection but want to investigate the behavior further.

In a comparison of thermography and optical tomography Mohr et al. [2020] investigated the suitability of the systems to detect artificial defects. A high-frequency IR camera and the OT were installed on the same printer, which allowed for the simultaneous monitoring of one buildjob. Additionally, the printer-integrated meltpool monitoring system was activated to capture data but was not used for detailed analysis as the proprietary data format limited data processing. All three systems were compared qualitatively to a post-process CT scan as shown in Figure 3.4. For the shown sample, all systems displayed apparent signal deviations. In comparison with other samples printed with standard parameters, the indications were less definitive and did not allow for a prediction based on only one sensor and the developed data analysis tools. According to the authors, the meltpool monitoring system showed the best qualitative correlation with the actual lack of fusion defects. They conclude that the sensor setup shows promising results but does not yet allow an automated evaluation. This is primarily due to the limited data processing and analysis. Hence, data fusion and enhanced data analysis were identified as important next steps.

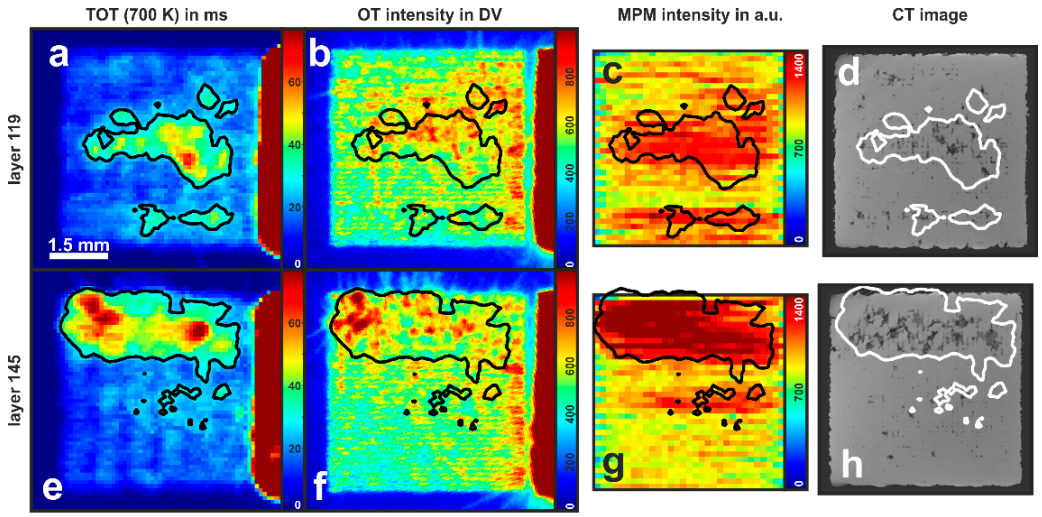


Figure 3.4 Comparison of the monitoring systems with CT. a)/ e) false-color plot of the time over temperature threshold (700 K) as captured by the IR camera. b)/ f) OT intensity as monitored by optical tomography. c)/ g) intensity of the meltpool monitoring system by SLM. d)/ h) CT image. The contour line is based on the time over temperature threshold of 30 $30\,µm$ [Mohr et al., 2020].

3.2.2 Machine Learning in Online Monitoring

In particular, the use of machine learning algorithms has shown great potential in recent years to enhance the analysis of monitoring data. In the following, the studies most relevant to this work are presented. For a more detailed review, the reader is referred to the literature [Razvi et al., 2019; Goh et al., 2020; Mahmoud et al., 2021].

In Baumgartl et al. [2020], the authors trained a neural network to classify false color images recorded by a thermographic camera during the LPBF process. The images either contained a delamination, splatter, or no anomaly. The system achieved an average balanced accuracy of 96.80%. Despite the already promising result, a couple of limitations to the work were identified. Firstly, the generalizability of the network has yet to be evaluated. Secondly, the study was limited to splatter and delaminations and a single geometry. Nevertheless, the potential of neural networks for classifying online monitoring data could be shown.

McGowan et al. [2022] also used a neural network to classify thermographic images. The authors used an off-axis pyrometer to monitor the meltpool. A post-process CT scan was performed to determine the porosity. Labeling the thermographic images into two classes (does/does not contain a pore) created a training dataset for the neural network classifier. A physics-informed loss function was introduced to the network training, which did not significantly improve the network performance. The accuracy for all investigated models was around 90%. Overall, the developed network performed worse than the PyroNet [Guo et al., 2020] network, which was chosen as a baseline as it followed a similar approach. The authors concluded that this is primarily due to hyperparameter tuning, data augmentation, and data partitioning. The paper summarizes the viability and need for further research in this area.

Gobert et al. [2018] trained a support vector machine (SVM) to detect defects based on gray value images. The images were captured by a high-resolution camera under eight different lightning scenarios per layer before and after recoating. The size and location of defects in the printed specimen were extracted using CT. To allow for a comparison of both data domains (online monitoring and CT), both data sets are registered to the original CAD file, allowing for a one-to-one mapping of coordinates. By this mapping, the recorded images can be divided according to the existence of a defect in the corresponding CT layer. The SVM achieved an accuracy between 63% and 73% when trained on a single lightning scenario. When incorporating all eight light sources into one ensemble classifier, the accuracy increased to 85%. This shows the importance of data fusion of different sensors. The authors conclude that in-situ layerwise imaging holds great potential, particularly in combination with machine learning approaches.

Scime et al. [2020] developed a neural network to analyze images captured by different cameras on different printers. The work focuses on the model transferability between different sensor setups. The authors could show that transfer learning allows for the adoption of one architecture to different sensor setups. This facilitates the model training by reducing the required amount of ground truth data and computational effort. Furthermore, it shows the feasibility of neural networks for machine transferability.

Yuan et al. [2018] investigated the suitability of a CNN to predict the laser track width, its standard deviation, and its continuity over time using an in-situ captured video. Ex-situ measurements of the laser track were used as ground truth labels for the CNN training. While the model already performed well on a small dataset, the authors emphasize that the prediction performance might increase with additional data and sensors.

Other approaches focus on the detection or classification of process signatures. The signatures are associated with a higher likelihood of producing defects. Therefore, the aim is to allow for a monitoring of process deviations as a possible indicator for defects. Scime and Beuth [2019] and Betts and Cola [2018a] show examples for this approach. While increasing the process understanding and possibly enhancing the process stability, classifying process deviations might lead to a drastic increase in falsely scraped parts, as a process deviation does not necessarily lead to a defect.

In summary, the current state of the art shows promising approaches for sensor setups and parameter monitoring. Different studies have indicated that monitoring systems can be used to characterize the process stability. There are no clear guidelines or recommendations concerning monitoring setups. Instead, many different sensors and evaluation techniques have been proposed in the literature. Furthermore, a

physical model linking monitoring signatures and post-process anomalies (e.g., pores) has yet to be established. Hence, the targeted research objective of using artificial intelligence to statistically investigate this link might provide new insights and understanding of the process and its monitoring. In particular, the systematic correlation and subsequent quantification of computed tomography and monitoring data for a statistically relevant number of samples should add to the current research within this field and allow for an improved automatic monitoring of the AM process.

4 Methods

As pointed out previously, this work focuses on the detection of defects based on a novel CNN approach. It implements a data processing pipeline to utilize existing sensor setups as various sensor systems already exist on the market. The following presents the CNN pipeline developed within this study in detail. The pipeline is illustrated in Figure 4.1. It takes the OM and CT data as inputs and trains an OM CNN to evaluate the monitoring data automatically. The CT data is used as the reference as it details the location and size of defects in the printed part.

The following chapters detail the individual steps developed within this pipeline. For the OM data, these include the production of the AM specimens with artificial defects (Section 4.1) and the construction of 3D volumes containing the OM data (Section 4.3.1). For the CT data, a registration procedure to align the OM and CT data is developed (Section 4.3.2) as well as a labeling method to (automatically) detect defects in the CT scan (Section 4.3.3 and Section 4.5). The core of the pipeline is the two CNNs developed to evaluate the CT and OM data automatically. Section 4.5 and Section 4.6 detail the respective CNNs. Besides the general feasibility, a special focus is placed on the experimental investigation of possible limitations. Therefore, different models are presented to evaluate the proposed approach's performance, transferability, and explainability. The motivation for each investigation is presented in Section 4.5 and Section 4.6 with the results shown in Chapter 5.

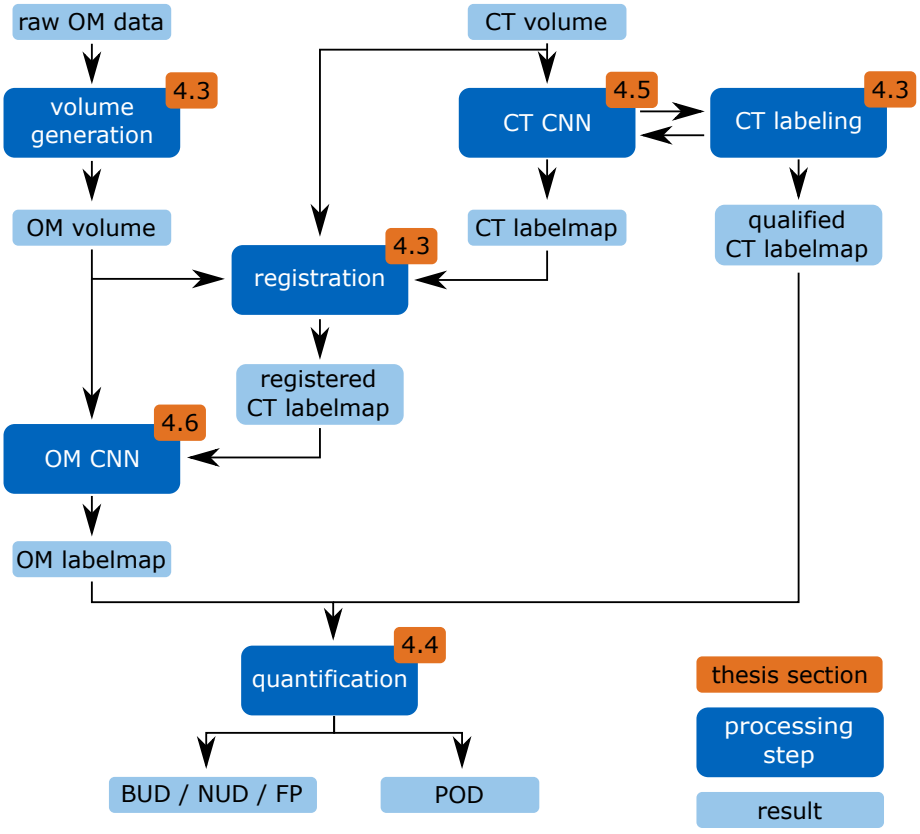


Figure 4.1 Overview of the developed pipeline for correlating CT and OM data by training convolutional neural networks.

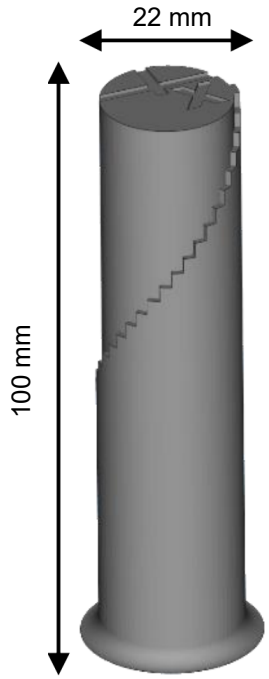


Figure 4.2 3D view of the specimen geometry used in this study. The specimens are approximately 10 cm in height, 22 mm in diameter and have an encircling staircase of 36 steps.

4.1 Sample Manufacturing

For this study, a total of 156 cylindrical samples and eight more complex geometries are printed. All samples are produced on a *GE ConceptLaser M2 Dual Laser* machine with Ti-6Al-4V powder. Figure 4.2 shows the geometry of the cylindrical samples. They consist of a solid cylinder (20 mm diameter) in the middle, which is encircled by a staircase and has a height of approximately 10 cm. The rough geometry is adapted from Gobert et al. [2018]. The staircase consists of 36 steps so that one step spans 10° of the outer diameter of the cylinder. Each step has a width of 1 mm and height depending on the height of the entire sample. Each step offers multiple well-defined spots that can be used to align the volume in a defined orientation. At the same time, the number and severity of CT artifacts is limited by the cylindrical shape. Therefore, the geometry allows for a good trade-off between a low CT artifact generation and a well-performing registration. A number on the top of the sample uniquely identifies the samples. The samples are produced in five buildjobs. Each buildjob is manufactured with different process parameters in order to provoke different sample properties. The buildjobs are introduced in the following and summarized in Table 4.1. Figures 4.3 and 4.4 illustrate the buildplate of the printer with the specimens with different process parameters marked by color. Additionally, the two inner red points mark the spots of direct laser reflection. At these locations, the laser hits the metal powder perpendicular, which might lead to a direct reflection of the laser beam back into the optical system, possibly affecting the monitoring system.

Buildjob A

Buildjob A contains reference samples printed with standard parameters, which are known to produce high-quality specimens with a high density and a low number of anomalies. They are marked in blue in Figure 4.3. The samples are classified as anomaly-free in the subsequent CT scan described in Section 4.2.2. The remaining specimens in Buildjob A (marked in gray) are printed with sub-optimal process parameters. In previous studies, this process parameter (i.e., the decrease in laser power) was found to provoke lack of fusion anomalies in the specimen. Each specimen is printed with constant process parameters. Therefore, the location and size of anomalies are not specified by altering the process at pre-defined locations but rather depend on statistical phenomena. This is an important difference, as previous studies

have primarily focused on the classification and localization of process anomalies, which are induced by altering the process parameters locally. While these previous approaches allow for the detection of process deviations, they are very limited in detecting actual anomalies in the printed part. The detection of actual anomalies in the part requires their statistical appearance, as local artificial process parameter changes would influence the sensor data systematically. Hence, the process parameter is chosen to provoke lack of fusion anomalies in the specimen while allowing for a stable building process and a stable specimen. The analysis of the lack of fusion anomalies is presented in Section 5.1.

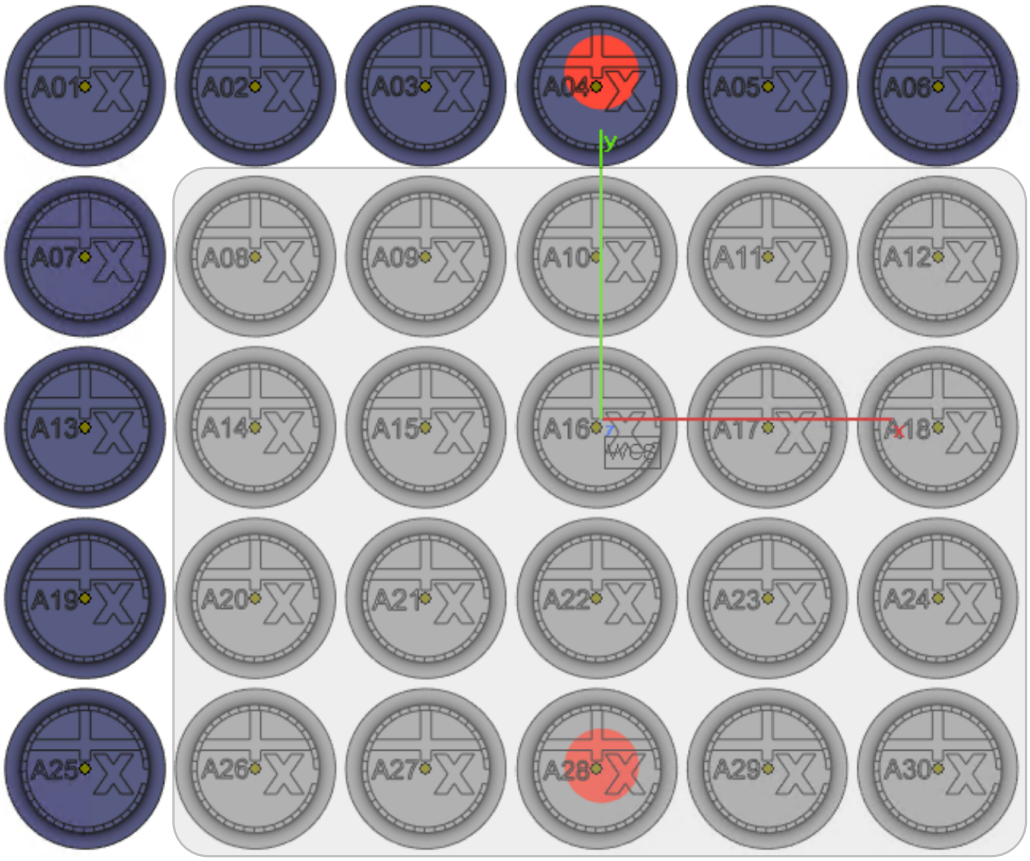


Figure 4.3 Buildplate of the Buildjob A with 30 specimens. Specimens marked in blue are printed with standard parameters. Specimens marked in grey are printed with parameters prone to produce anomalies in the specimen.

Buildjob B

Buildjob B also contains reference samples (B01, B07, and B13)(green region in Figure 4.4). These samples are printed with 100% of the standard focus. The other samples are printed with parameters prone to produce keyhole anomalies. In this case, two process parameters are altered. For the top 18 specimens (blue region in Figure 4.4), the laser focus is decreased. For samples B02, B03, B08, B09, B14, and B15, the laser focus is decreased to approximately 45%. For the samples B04-B06, B10-B12, and B16-B18, the focus is further decreased to approximately 40%. In this case, the decrease in focus means focusing the same laser energy on a smaller region. Hence, the energy density at these points is higher, leading to a higher energy input on specific points. At the same time, the region directly influenced by the laser is smaller. This may lead to a higher probability of keyhole creation.

For samples B19-B30, the laser focus is kept at the standard parameter, but the skywriting lead time is altered (orange region in Figure 4.4). As introduced in section 2.1.1, skywriting compensates for the decrease in laser speed at the end and the beginning of a scanning line. Due to the inertia of the optical system, the scanner is slower at the beginning of a scanning line as it is still in the acceleration phase. If the laser power is kept at a constant high during this period, this leads to a higher energy input as the laser

is focused on one point for longer. To combat this behavior, the laser is turned off briefly while the scanner turns to the next scanning line and accelerates to the desired speed [Mancisidor et al., 2016]. This time window is defined as skywriting lead time. Therefore, a decrease in skywriting lead time may increase energy input at points where the scanner system has to accelerate or decelerate.

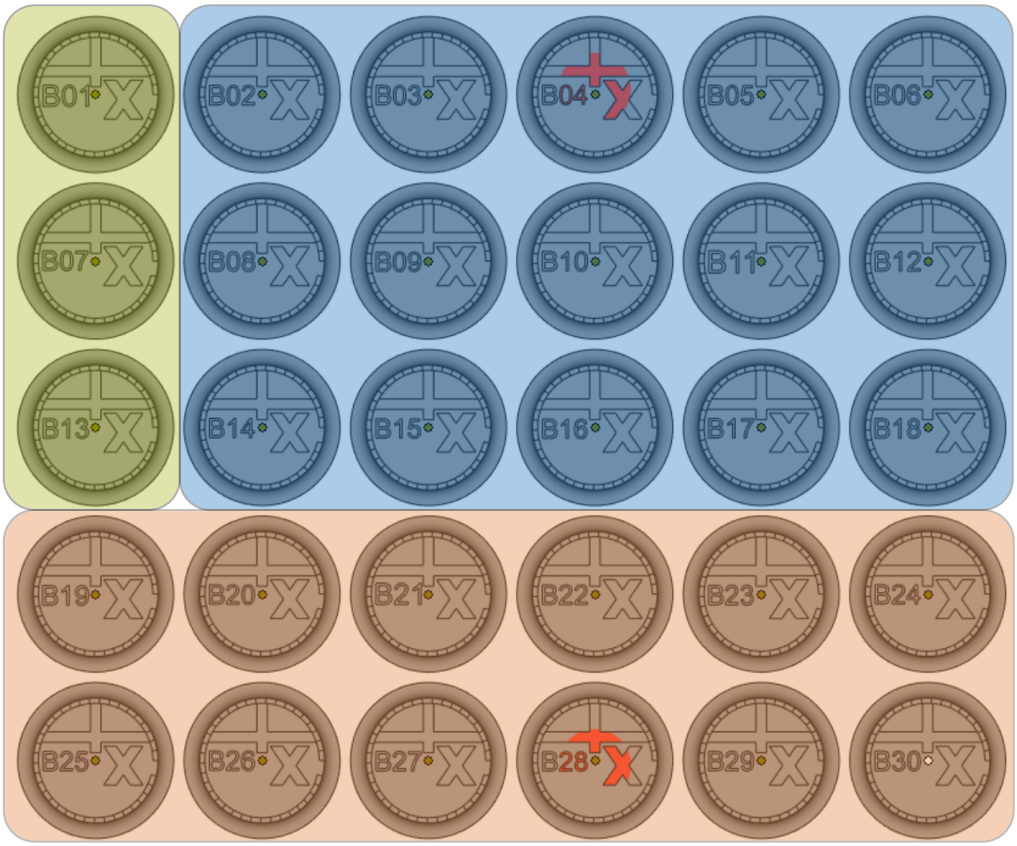


Figure 4.4 Buildplate of the Buildjob B with 30 specimens. The specimens are printed with varying focus (blue region) or decreased skywriting lead time (orange region). Additionally, reference specimens printed with standard parameters are marked in green.

Buildjob C

In contrast to Buildjob A and B, Buildjob C is printed with reference parameters. Hence, the specimens are expected to have no lack of fusion or keyhole anomalies. Instead, the powder is contaminated with tungsten powder to simulate inclusions in the specimen. Tungsten might occur as inclusion as tungsten electrodes are used in the powder production of Ti-6Al-4V [Platacis et al., 2019; Withers and Loutfy, 2014]. For this purpose, 2% (weight percent) pure tungsten powder is added and mixed with the Ti-6Al-4V powder. The tungsten particles had a size distribution of: $D_{10} = 52$µm, $D_{50} = 61$µm and $D_{90} = 73$µm. As a result, all specimens in Buildjob C contain inclusions over the entire specimen. Due to the small size distribution of the powder, no tungsten particles with a relevant diameter are observed within the specimen. Therefore, the Buildjob C is only used to train the CT CNN. No specimens are used for the training of the OM CNN as the created inclusions are considered not relevant for the detection via online monitoring.

Buildjob D

Buildjob D combines multiple anomaly types. Similarly to Buildjob C, tungsten powder is added to the titanium powder. In this case, the amount of tungsten powder is drastically reduced, and the powder is added without mixing the powder afterward. Hence, the tungsten inclusions are concentrated in two distinct layer regions around 2 cm and 4 cm sample height.

Table 4.1 Summary of the buildjobs printed in the project with the different process parameters used to provoke anomalies in the specimens.

Buildjob	Process Parameter	Comment
Buildjob A	increased scan speed	parameter prone to create lack of fusion anomalies
Buildjob B	decreased focus/ skywriting delay	parameter prone to create keyhole anomalies due to higher energy input by smaller focus (defocus) or slower laser (skywriting)
Buildjob C	tungsten inclusions	higher density inclusions with size around 60 µm
Buildjob D	mix of A, B, C	combination of specimens printed with different process parameters and inclusions

The specimens D01-D06 and D26-D31 are printed with the same lack of fusion parameters as the samples in Buildjob A. Therefore, they combine possible lack of fusion anomalies with inclusions. The specimens D08-D13 and D32-D36 are printed with standard parameters and should, therefore, not contain relevant pore (keyhole/ lack of fusion) anomalies. The specimens D14-D25 are produced with a reduced focus of approximately 40%, similar to the specimens in Buildjob B.

Additional Buildjobs

Besides the buildjobs described above, additional specimens are produced during this thesis. For clarity and readability reasons, those are not described in more detail here as they did not contribute in a relevant way to the discussed results. Instead, they are listed and described in short in Appendix 1 and Appendix 2.

4.2 Data Generation

For the successful implementation of artificial intelligence algorithms, adequate data gathering and processing is essential. In this thesis, this includes the data gathered during the AM process (online monitoring data) and the CT data gained in a post-process CT scan. The correct alignment of both datasets and the corresponding pre-processing steps are described in Section 4.3.2.

4.2.1 Online Monitoring

As described in Section 3.2, there exists a variety of different monitoring solutions for LPBF machines. For this research, the *Sigma Additive Solution* Meltpool system *PrintRite3D* is installed on the *GE Concept Laser M2* machine as sensor hardware. The sensor system consists of three photodiodes installed in parallel to the laser source in the optical path, as shown in Figure 4.5. The laser beam passes through a semi-transparent mirror, is redirected by the scanner, and reaches the powder-coated buildplate. Subsequently, it melts the powder and hence generates a meltpool. The meltpool itself emits radiation, part of which then is passed back through the scanner and semi-transparent mirror. The radiation is split by a beam splitter and filtered to process specific wavelengths per photodiode.

One photodiode captures a wide band of wavelengths and is named *Thermal Emission Density (TED)* by SGLB [Megahed et al., 2019]. The other two photodiodes are sensitive to a narrow band of wavelengths, which together are supposed to be representative of the temperature of the meltpool. The ratio of both photodiode signals is titled *Thermal Emission Planck (TEP)* and should correlate to the temperature of the meltpool [Betts and Jacquemetton, 2018, 2019]. Figure 4.6 shows the different wavelengths exemplary.

In the scope of this study, the processing of the voltage measurements defined by *Sigma Additive Solution* showed potential for improvement as the provided output contains only a limited amount of information

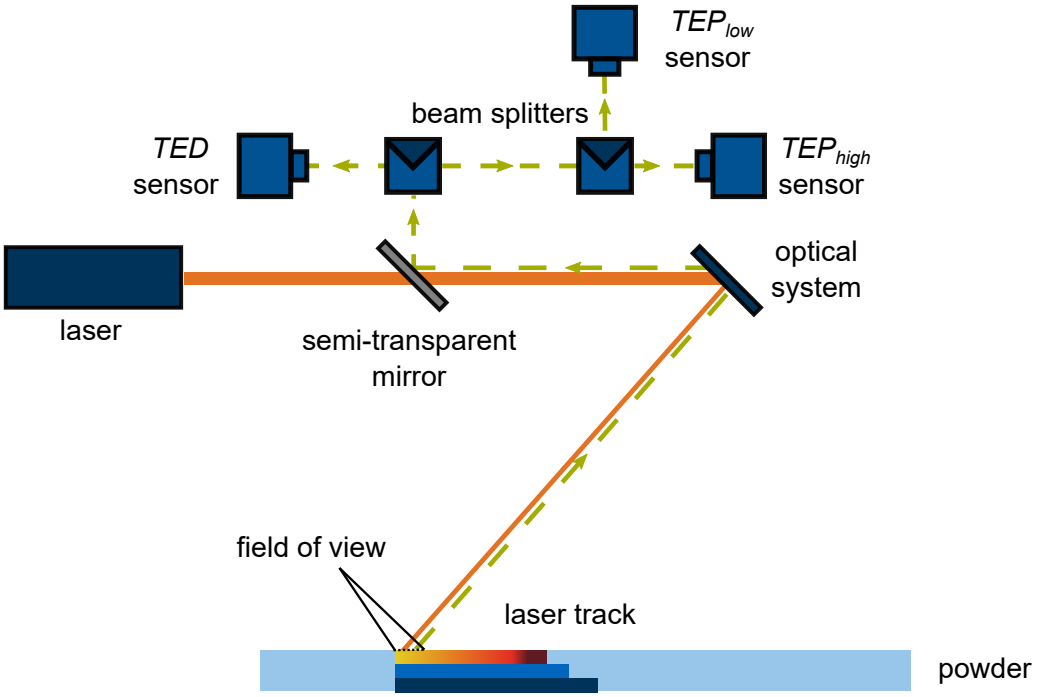


Figure 4.5 Schematic demonstration of the implemented on-axis online monitoring system by *Sigma Additive Solutions*.

for the training of AI algorithms. Nevertheless, the hardware setup itself is identified to gather relevant information on the meltpool. Therefore, the raw data produced by the photodiodes is used as input for further investigation. Following the *Sigma Additive Solution* nomenclature, the three channels provided by the hardware setup are referred to as TED, TEP_{high}, and TEP_{low}. Where $TEP_{high}$and TEP_{low}correspond to the signals captured by the photodiodes with the narrow higher and lower bandwidth, respectively.

The photodiodes measure the radiation intensity at a sample rate of 200 kHz ($TEP_{low}$and TEP_{high}) and 100 kHz (TED). Additionally, the system tracks the relevant machine parameters at a rate of 100 kHz. Those are the set *Laser Drive* signal, the set *Laser Power* signal, and the set *x* and *y* signal. The data is logged per layer and laser and is provided as a *hdf5* file. Table 4.2 lists and explains the content of the hdf5 file.

As pointed out, the SGLB setup is only used as a hardware sensor setup. The data processing and data fusion necessary for the analysis are developed in the scope of the project and are introduced in Section 4.3.1 and Section 4.3.2. This should allow for an easy transfer of the investigated approach to different monitoring sensor setups.

4.2.2 Computed Tomography

The CT scans are conducted by qualified inspectors of the *Testia GmbH* in Ottobrunn. The CT machine used for all scans is a *Diondo d2 300 kV*. The cylindrical specimens are put in a specifically designed specimen holder shown in Figure 4.7. The holder is made of a polymer, which allows for a low absorption of X-rays compared to the titanium samples. Hence, its influence on the CT scan can be regarded as marginal. The mount can be loaded with four samples stacked on top of each other for a series scan. Therefore, each scan only contains one specimen, allowing for a high-quality scan and an easy separation of the specimens. The scan is performed with an acceleration voltage of 250 kV and a current of 120 µA, which results in a power of 30 W. The X-ray beam is pre-filtered by a 2.5 mm copper filter to reduce beam hardening artifacts. The voxel size of the reconstructed volume is set to 30 µm. This resulted in a reconstructed volume of 2208 voxel x 2208 voxel x 4016 voxel. The reconstruction is performed with an integrated filter by Diondo to reduce ring artifacts. The resulting volume is saved as a raw image stack where the entire volume is sliced in the xy-plane, and each slice is saved as a raw image. For the cylindri-

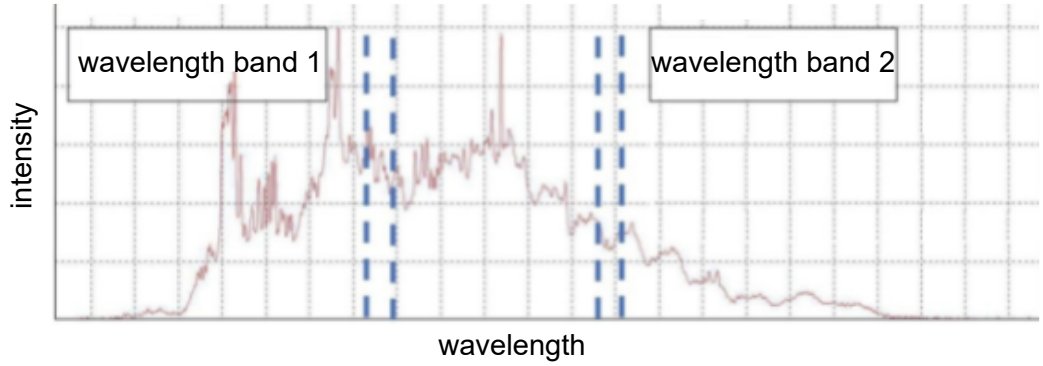


Figure 4.6 Exemplary intensity curve of the meltpool emission over the wavelength. Illustrative bandwidths for the narrow bandwidth photodiodes are marked in blue. Wavelength Band 1 corresponds to TEP_{low} and Wavelength Band 2 corresponds to TEP_{high} (adapted from [Betts and Jacquemetton, 2018], text quality improved due to low resolution in original paper).

Table 4.2 Content of the hdf5 file provided by the *Sigma Additive Solution* system. The value ranges are machine-specific and are extracted empirically during this research.

Parameter	Value Range [V]	Comment
Laser Drive	[0, 5]	Binary signal to specify (no-)movement of laser.
Laser Power	0 – 3.5	Continuous signal to control the laser power. Depends on the defined printing parameters.
TEP_{high}	0 – 4.3	Continuous signal of the TEP_{high} photodiode.
TEP_{low}	0 – 4.3	Continuous signal of the TEP_{low} photodiode.
TED	0 – 10	Continuous signal of the TED photodiode.
X-Position	-5 – 5	Continuous signal to control the X-Position of the laser on the build platform by steering the scanner.
Y-Position	-8.2 – 7.9	Continuous signal to control the Y-Position of the laser on the build platform by steering the scanner.

cal specimens, this resulted in 4016 raw images. Figure 4.8 shows an example of a reconstructed volume (contrast adjusted manually for better visualization) and a corresponding raw image slice. In Figure 4.8b, the C-shaped specimen holder can be seen in light gray as it has a lower density than the titanium specimen but a higher density than the surrounding air.

For the subsequent processing steps, the individual raw images have to be combined into a 3D volume to allow for an efficient registration and analysis. The chosen 3D image format for this purpose is the *Neuroimaging Informatics Technology Initiative* (nifti) format [Cox et al., 2004]. The nifti format is one of the major data formats in medical imaging and consists of the actual image volume and a header [Larobina and Murino, 2014]. The header contains supplementary information about the volume, such as the orientation of the volume, the data type the voxel information is saved in (e.g., uint16), and the physical size of a voxel (e.g., one voxel corresponds to $30\,\mu m$). The nifti format is widely adopted by researchers, allowing easy integration in existing frameworks and programs [Li et al., 2016].

The conversion from raw image stack to nifti is implemented in Python using SimpleITK [Lowekamp et al., 2013] In the first step, the raw image is imported as a 2D array (uint16). Subsequently, the 2D arrays are stacked in z-direction, resulting in a 3D array of size 2208 x 2208 x 4016. This array is then converted to a SimpleITK array, which can store the additional information for the nifti header. As seen in Figure 4.8b

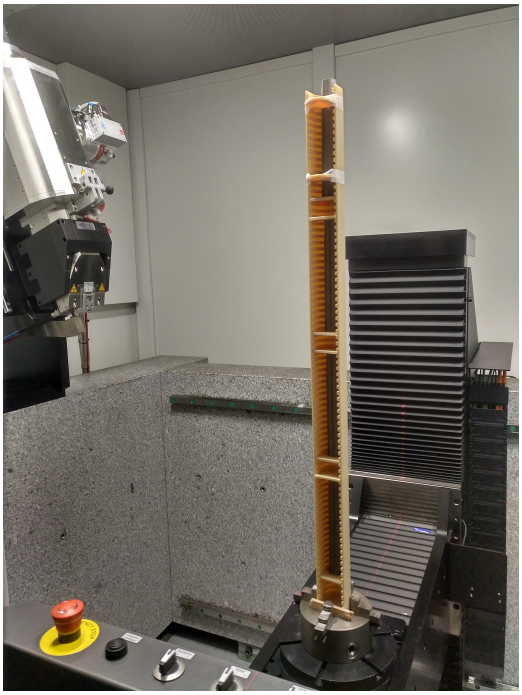
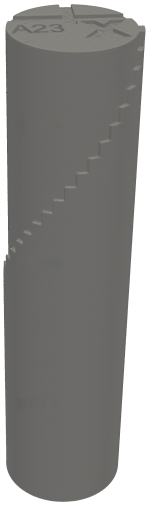
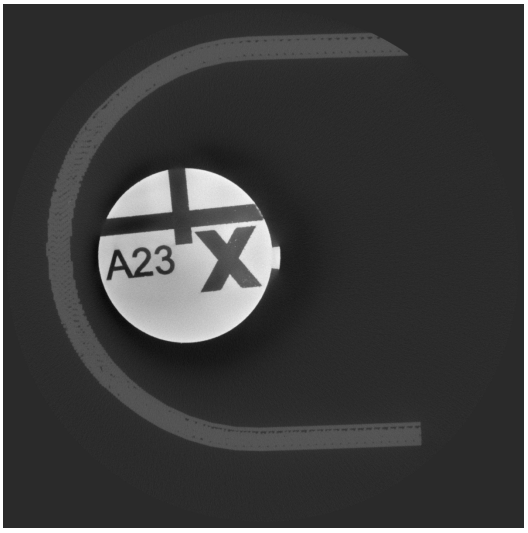


Figure 4.7 CT scan setup for scanning the cylindrical samples with the specimen holder in the center fitted with four specimens. The X-ray tube is visible on the left while the detector is on the right-hand side out of frame.

the volume contains a large portion of irrelevant information as the specimen makes up only around one-quarter of the image. Hence, the volume is cropped to the relevant area containing the specimen. For this, a binary map of the sample is created by thresholding the volume. The threshold that best separates the specimen and the background is determined by Otsu's threshold [Otsu, 1979]. Subsequently, the volume is cropped to the bounding box of the specimen mask. Here, the specimen is identified in the binary mask as the largest connected component, as multiple smaller objects in the image might also be above the Otsu threshold. As a conservative approach, a safety padding of 50 voxels in each direction is added to the bounding box. After the cropping operation, the volume size could be reduced to around 850 x 850 x 3300 voxel, reducing the file size from around 18 GB to approximately 4 GB for the cylindrical specimens [Permadi, 2021].



(a) Reconstructed CT volume with manually adjusted contrast for better visualization. In particular, the surrounding specimen holder is removed by manual adjustment.



(b) Visualization of a raw image as it is reconstructed and stored by the Diondo software. The specimen is clearly visible in the middle of the image, surrounded by the C-shaped polymer specimen holder.

Figure 4.8 Reconstructed CT volume and a corresponding raw image slice.

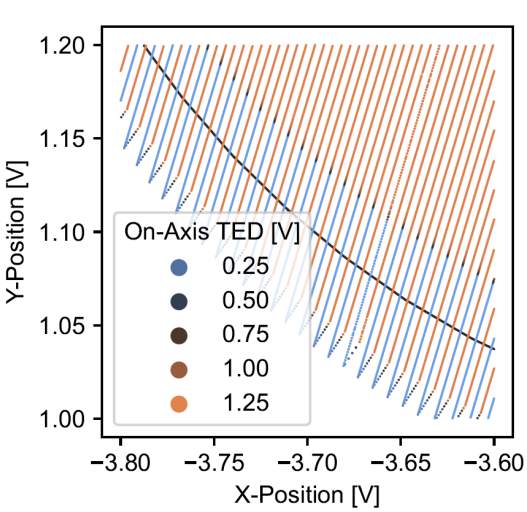
4.3 Data Pre-Processing

4.3.1 3D Volume Generation Online Monitoring

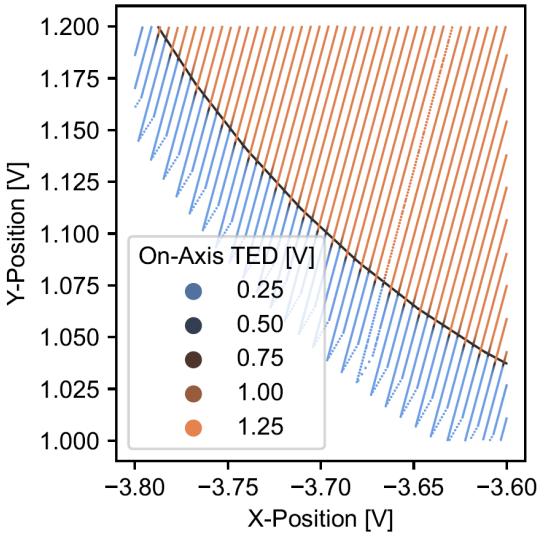
The hdf5 file, as described in Section 4.2.1, is taken as input for the following pre-processing steps. Due to imperfections in the sampling process, the channels in the hdf5 are sampled with different time delays. To synchronize the measurements with each other, the time delay is corrected by shifting the channels respectively. The photodiode signals (TEP_{low}, TEP_{high}, and TED) are taken as reference as they are the relevant measurements for the later analysis. To obtain the correct shifting factor, the time series of the individual channels are plotted as seen in Figure 4.9. Then, the signal is analyzed manually by inspecting the signal behavior at the edges of the sample. The purple line in Figure 4.9a shows the contour of the specimen. Hence, outside of this line, the laser is expected to be turned off as no powder is to be melted there. Therefore, at these points, the TED signal is expected to be zero. By shifting the x- and y-position time series, this is achieved as shown by Figure 4.9b. The x- and y-position signals had to be shifted by 43 timesteps to be in sync with the photodiode signals. Similarly, the laser drive is shifted by 104 timesteps. The values are determined empirically by manual inspection of the xy-plot as shown in Figure 4.9 as well as analyzing the time series data of the different signals.

The adapted time series data is used to construct a 3D volume containing the relevant information of the photodiode signals but represented in a spatial reference frame similar to the CT data introduced in Section 4.2.2. For this, the x- and y-information stored in the hdf5 file is used to assign the corresponding TED, TEP_{low}, and TEP_{high}signals to the related coordinates in a 3D volume. Hence, the temporal relation of the data is lost in return for a spatial relation of the individual data points. This allows for a correlation between the CT data and the OM data. In this study, a variety of different transformations is developed.

The general idea of all the mapping processes is the conversion from a temporal to a spatial reference frame with minimal loss of information. Each time step in the pre-processed time series contains information about the location of the laser at this point in time combined with the corresponding photodiode signals and the process parameters as described in Section 4.2.1. As for the CT data, the spatial representation is implemented on a voxel basis.



(a) Visualization of the TED signal without a shift in the x- and y-position signal. The region in which the signal drops significantly (blue points) does not correspond with the actual region in which the laser is expected to be turned off as it is outside of the sample (black line shows the contour of the sample).



(b) Visualization of the TED signal with a shift of 43 timesteps in the x- and y-position signal. The captured TED signal corresponds well with the expected result as it drops significantly for regions outside of the sample.

Figure 4.9 Comparison of the TED signal before (a) and after (b) the signal shift to correct the imperfect signal sampling.

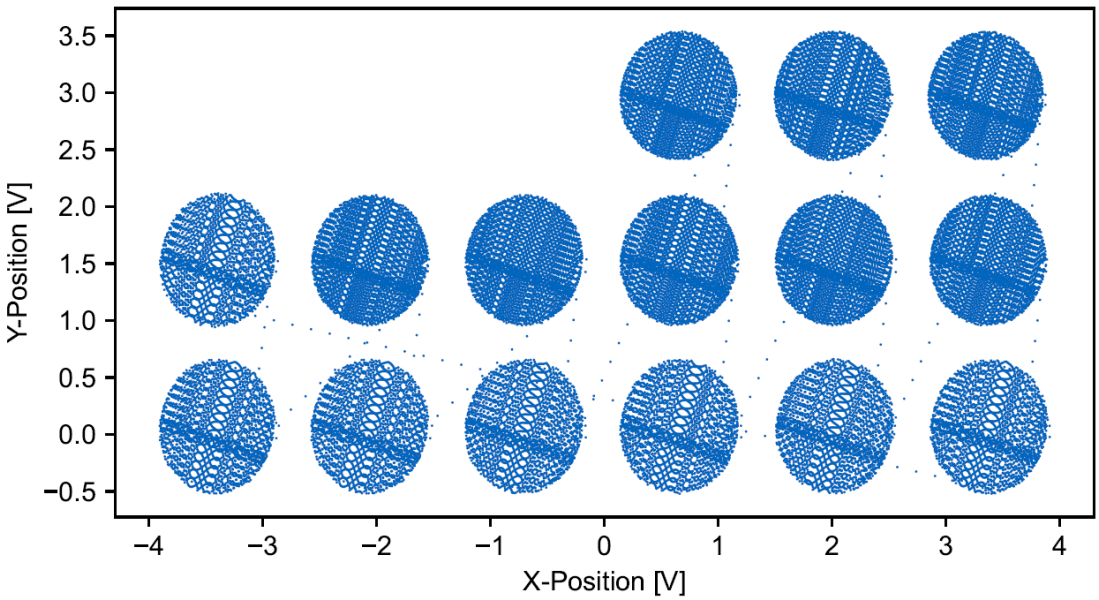


Figure 4.10 Plot of the Laser Drive signal of one hdf5 file containing the time signal of one optical system for one entire layer. The patterns visible inside of the samples are an aliasing artifact, as only every 100th point is plotted for visibility reasons.

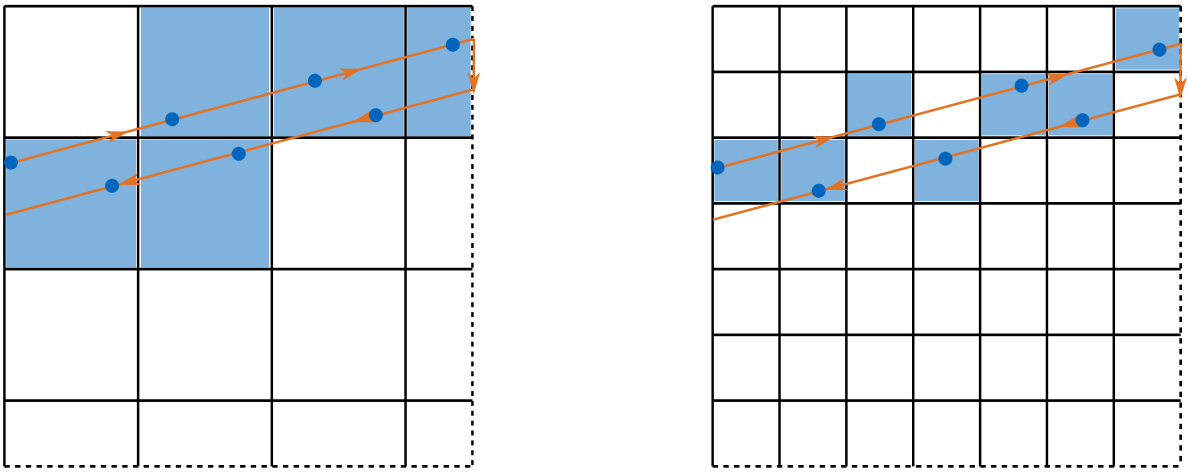


Figure 4.11 Visualization of the mapping process. The buildplate is divided into individual cells represented by a 2D array. The cells are well-defined by their borders. Each time series data point is mapped onto this grid by its corresponding xy-coordinates. Hence, one grid cell may contain no, one, or multiple data points.

One hdf5 file contains the information of both optical systems for an entire layer of the build platform. An example is illustrated in Figure 4.10 which shows the Laser Drive signal plotted at the corresponding x and y coordinate. In the first step, the hdf5 file with the data for the two optical systems is imported.

Secondly, as the x-/ y-signal is encoded as a voltage signal, the voltage has to be transformed to a spatial coordinate on the buildplate. This transfer function has to be computed per machine by printing a reference buildjob. The reference buildjob consists of reference points at pre-defined coordinates. By capturing the voltage signals at these points and computing their relation to the preset xy-coordinates, the transfer function can be estimated numerically. Additionally, to the direct voltage-to-coordinate conversion, a distortion correction is implemented in this step. The distortion correction can be estimated by the same reference buildjob and aims to compensate for the distortion introduced into the system by the scanner aperture.

Thirdly, the desired array grid is defined. This step defines the xy-resolution of the 3D volume as it determines the size of one grid cell in the array. For example, if the entire buildplate has a size of 245 mm and the array is defined as 1000 x 1000, the resulting size of one grid cell (and hence the resolution) would be 245/1000 mm. Therefore, every point on the buildplate is assigned to one specific grid cell. On the other hand, one grid cell does not necessarily contain only one data point, as the distance between two data points may be smaller than the grid size. This is illustrated in Figure 4.11. The maximal resolution is limited by the sampling rate of the sensors. In the case of multiple values per cell, a decision on the resulting cell value has to be taken with the objective of minimal information loss, e.g., the cell value could be set to the mean value of all data points within this cell. As the process itself and the monitoring principle do not allow for a definitive answer on the relevance of extremes or how anomalies might be represented in the data, a data-driven approach is chosen. By applying different pre-processing strategies and observing their influence on the result, the most relevant information should be preserved, and a conclusion on the relevant data features might be possible. Different strategies are investigated during this thesis and are presented in the following.

Maximum-Minimum Mapping

In the case of the maximum-minimum mapping, two values are selected per cell. One array is created by selecting the maximum value per cell. Another array contains the minimum value per cell. Both arrays are stored in individual 3D volumes in nifti format. This allows for the retention of both extremes but doubles the data size. The focus on extremes might facilitate the detection of anomalies, as anomalies might be represented as outliers in the signal. Additionally, the maximum-minimum mapping might allow for a more

thorough analysis of relevant signal signatures in a later step. On the other hand, the integration of two volumes per channel might be redundant or even hinder the training of CNNs.

Largest Deviation Mapping

To reduce the data size and avoid a possible redundancy introduced by storing the minimum and maximum value, the largest deviation mapping is introduced. In this case, the largest deviation from the mean value of all data points within the cell is chosen. For this, the mean value is computed, and its difference to every data point is calculated. The value with the largest difference is assigned to the cell. This allows for the retention of the strongest extreme but does omit the weaker extreme as well as the information about other data points in the cell.

Radial Distance Mapping

The motivation for the radial distance mapping is the spatial extension of the meltpool or, rather, the area influenced by the laser. In the case of the maximum-minimum and largest deviation mapping, the measurement is assigned to a discrete point in space without an extension. In reality, the laser might affect neighboring regions as well, and hence, the measurement taken at this point is extended to neighboring regions. The influence of one measurement on surrounding regions is expected to decrease with distance. Therefore, the radial distance mapping takes the distance-weighted sum of all data points within 3x3 grid cells in the xy-plane. The region taken into consideration for the mapping is established empirically.

Track Mapping

In contrast to the previously introduced mapping strategies, track mapping incorporates, in a limited manner, the temporal information in the mapping process. For this, the temporal context of the individual data points is considered. If, for example, the points within one grid cell are melted or influenced by the laser multiple times, the chronological order of the individual measurements might be relevant. In the case of the track mapping, the latest measurement is assigned to the entire cell. Therefore, if a cell already contains a value, this value is replaced by the new value if the corresponding measurement falls within the grid cell. This approach might also allow for the detection of systematic shifts over time, e.g., if the laser power decreases over time.

3D Volume Generation

Independent of the mapping method, the resulting arrays are further processed by stacking the created 2D arrays in z-direction. The resulting 3D volumes contain the corresponding information for all slices of the entire buildplate. In order to extract the information per sample, the buildplate volume is split into subvolumes for each sample. This approach is similar to the pre-processing of the CT volume in Section 4.2.2. For this, a mask volume is created by binarizing the 3D volume. As all pixels outside of a specimen are zero, the binarization can be achieved by a simple threshold of one. Subsequently, the individual volumes in the binarized masked are uniquely labeled, and the corresponding bounding box is computed. By cropping the 3D volume to the computed bounding box (plus a padding as a safety cushion around the actual volume), a 3D volume of a single sample is obtained. This volume is converted to the nifti format and saved with the associated image properties.

In contrast to the CT volume, the voxel size might be anisotropic as the resolution in the xy-plane is defined by the grid size set introduced above, while the resolution in the z-direction is defined by the layer thickness in the manufacturing process. Therefore, for the samples in this thesis, the voxel size in the z-direction is fixed to $60\,\mu m$ as this is the thickness of one layer of powder in the printing process. The voxel size in the xy-direction can be refined by adjusting the grid size applied to the buildplate and is a trade-off between file size and information value.

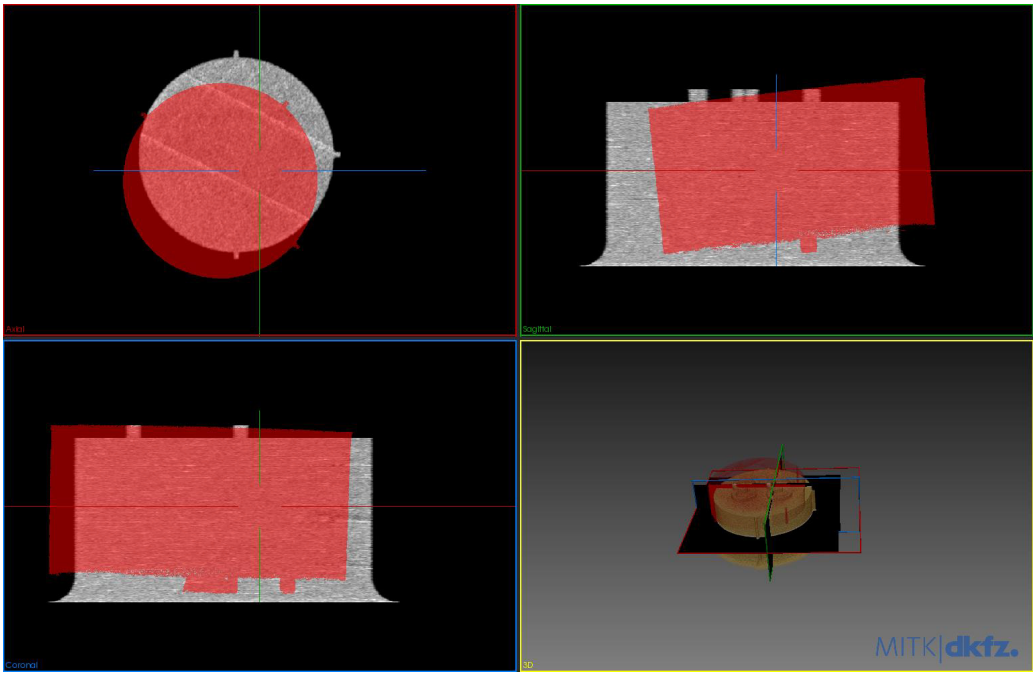


Figure 4.12 The CT volume (red) and the OM volume (gray) are represented as individual 3D volumes of the same specimen. To allow for a comparison and correlation of both datasets, they have to be registered to a common reference frame.

4.3.2 Registration

In Section 4.2.2 and Section 4.3.1, both datasets, the CT data and the OM data, are transformed to 3D voxel volume as a common representation form. Both datasets are split so that each volume only contains one sample and are saved as individual niftis. As a next step, the two volumes are registered to a common reference frame so that both volumes are oriented the same way. This allows for a comparison of individual voxels and is mandatory for the spatial correlation of CT and OM data, as can be seen in Figure 4.12. The registration is performed in four main steps as they are investigated in Montijo Badeira [2020].

Initialization

As a first step, a rough initialization of the two volumes is performed. The input position of the two volumes in space is arbitrary. Therefore, the initialization is used to center the two volumes relative to each other. It uses the moments of both volumes based on the gray values. For this, the intensity-based center of mass of both volumes is computed and subsequently aligned. Here, the OM volume is set as a fixed image, whereas the CT volume is moved to align with the center of mass of the OM volume. The moments-based initialization is robust against differences in volume size as it computes the center of the volume based on the gray value distribution and not based on the size of the image volume. Additionally, the CT image is resampled to the OM image space, i.e., the voxel size, and hence, the resolution of the volume is set to the OM domain.

Coarse Manual Registration

In the second step, the initialized volumes are compared manually. Both volumes are centered with respect to each other, but they are not rotationally aligned (similar to Figure 4.12). The automatic registration with an optimizer is sensitive to the initial rotation of the volumes. Therefore, the registration process is facilitated by a coarse manual alignment. In particular, the initial rotation for the cylindrical specimens is crucial for the correct convergence of the automatic registration. As can be seen in Figure 4.2, the cylindrical specimen is symmetric to the xy-plane centered in the sample and axis-symmetric to the z-axis,

with the exception of the steps and the marks on top of the sample. As a majority of voxels of the sample belong to the symmetric portion of the sample, the registration optimizer has only limited information to uniquely identify the correct symmetry. For example, the difference between the marked side of the specimen being on the top versus the bottom is only defined by a small portion of voxels in comparison to the overall volume. As the optimizer aims to minimize the difference between both volumes, it might align the volumes according to this symmetry if initialized with a wrong initial rotation. The coarse manual registration can be transferred for all cylindrical samples as they are all scanned with the same specimen holder (Figure 4.7) and, therefore, are all oriented in a similar way. More complex specimens have to be aligned individually.

Refined Registration

The third step takes the coarse registration as input and refines the registration by numerical optimization. For this, a rigid 3D Euler transform is assumed as no significant deformation of the specimens is expected. The goal of the optimization is to minimize the geometrical difference between the two volumes. The difference is estimated by the Mattes mutual information similarity metric [Mattes et al., 2003]. It provides a measure of how much information can be inferred about a random variable by looking at another random variable. If the similarity of the two variables is high, one can infer a lot of information about the second variable by simply looking at the first variable. In contrast to simply comparing the raw intensity values of voxels, this approach is well suited for multi-modality registration, i.e., the registration of data obtained from different sensors [Johnson et al., 2015]. The similarity metric is minimized using gradient descent.

This registration step allows for a quick refinement of the overall registration, taking into account the entire volume. For a more detailed investigation of the used parameters and registration setup, the reader is referred to Montijo Badeira [2020].

Enhanced Registration for Top Layers

In the fourth step, only the top layers of the respective volumes are used for registration. This step is specific to the cylindrical specimen geometry. As shown in Figure 4.8a, the cylinders are marked unambiguous by a cross, a large X, and the sample ID on top of the sample. In order to further improve the registration, those marks are used by focusing on the top layers of each volume. This requires the two volumes to be aligned fairly well in height, as the selected layers in both volumes (CT and OM) should contain the introduced marks.

By limiting both volumes to a smaller subsection, the influence of those marks is expected to increase as their share of the overall volume increases. Hence, the similarity metric is driven more strongly by the distinctive features on the sample, forcing the optimizer to focus on the correct alignment of them. The specific number of layers to be used is specified manually per buildjob or, in rare cases, for individual specimens if the registration failed otherwise. After cropping the volumes to the specified number of layers, the registration is performed analogously to the Refined Registration above.

The quality of the overall registration workflow and its limitations are discussed in more detail in Montijo Badeira [2020]. The obtained registrations are subsequently checked manually and refined if needed. The resulting transforms are saved and can be used to apply the same transformation steps to other volumes. In particular, the transformation is used to register the label map of CT anomalies to the OM domain.

4.3.3 Data Labeling

The supervised training of CNNs requires a representative dataset of relevant anomalies and non-anomalous volumes, the so-called Ground Truth (GT). As the goal of the CNN segmentation analysis is the voxelwise segmentation of a 3D volume (intensity image), a GT label map for each volume is required. The label map has the same dimension as the image volume and labels voxels that belong to a corresponding anomaly. In the case of binary classification (anomaly/ no anomaly), the label map contains a 1 for voxels with an

anomaly. Otherwise, the voxel value is 0. An example of such a label map is shown in Figure 4.13. Figure 4.14 shows a 3D representation of a CT volume with anomalies marked by the corresponding label map. The quality of the created label map is essential for the training and testing of the CNN. While an accurate per voxel segmentation influences the segmentation accuracy, the ability of the CNN to generalize and transfer the training to new data depends on the diversity and number of anomalies in the training dataset.

As the manual labeling of a large amount of data is time and cost-intensive, the label maps for this thesis are created in an iterative manner. The process is visualized in Figure 4.15. Firstly, a total of nine label maps are created by the semi-automatic labeling process described below and in Figure 4.16. Secondly, the label maps are used to train the CT CNN described in Section 4.5. This CT CNN is then applied to the CT volumes of the other samples, hence producing label maps for those samples as well. A selection of these label maps is then evaluated by qualified inspectors as described in Section 4.3.4. Based on the qualified label maps, the CT CNN is improved by finetuning, and the resulting CT CNN is then used to produce refined label maps on new and existing CT volumes. The individual steps are described in the following, while the results of the respective CT CNN trainings are shown in Section 4.5 and Section 5.1.

Semi-Automated Segmentation Workflow

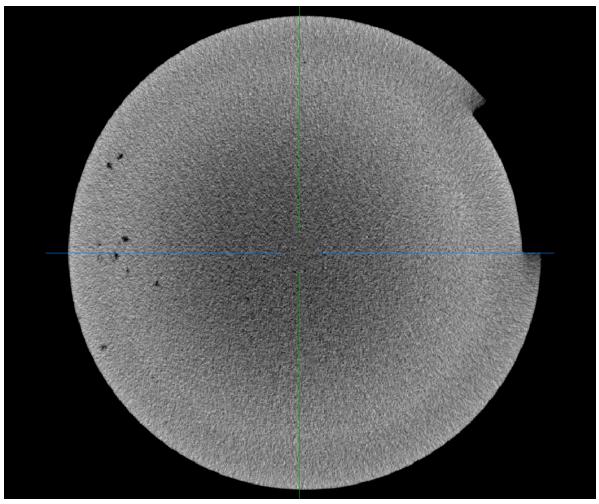
In the first step, a semi-automated workflow is developed and implemented in Python and SimpleITK. Figure 4.16 shows the corresponding flowchart. The workflow takes the unprocessed CT volume as input. Firstly, the specimen and the surrounding air are separated by applying Otsu's thresholding algorithm [Otsu, 1979]. This split is not perfect due to the low contrast between air and parts of the specimen, i.e., porous regions in the specimen. Therefore, possible holes in the specimen mask are closed by Binary Morphological Closing, and the mask is shrunk to exclude the surface of the sample. Subsequently, the largest component detected in the image (the specimen) is selected by the Connected Component Analysis (CC) and the Label Shape Statistics filter (LSS). The bounding box of this mask is used to crop the CT volume to the relevant region containing the specimen. Depending on CT artifacts, the top and bottom of the specimen are cropped additionally by human interaction to exclude regions with many falsely labeled anomalies.

In parallel, the actual anomalies within the specimen are segmented. For this purpose, a Median Smoothing filter is applied to the original intensity image to reduce the image noise. Subsequently, "small" darker regions in the intensity image are extracted by applying a Black Top Hat filter. The choice of kernel size for the filter defines which regions can be regarded as "small" and is set manually for the image. The filtered image is then binarized by applying Otsu thresholding combined with manual finetuning per image. The resulting binary label map is masked by the previously described specimen mask, removing all falsely detected regions outside of the specimen. In the final step, anomalies with a maximum elongation smaller than a defined threshold are removed in order to clean the label map of anomalies consisting of only a few voxels. As pointed out, this is a semi-automated process and requires interaction by a human non-expert. The different parameters (e.g., filter kernel sizes, thresholds) are optimized on a per-sample basis. Similarly, regions prone to falsely labeled anomalies are also excluded in a manual step. [Holtmann et al., 2021b; Permadi, 2021]

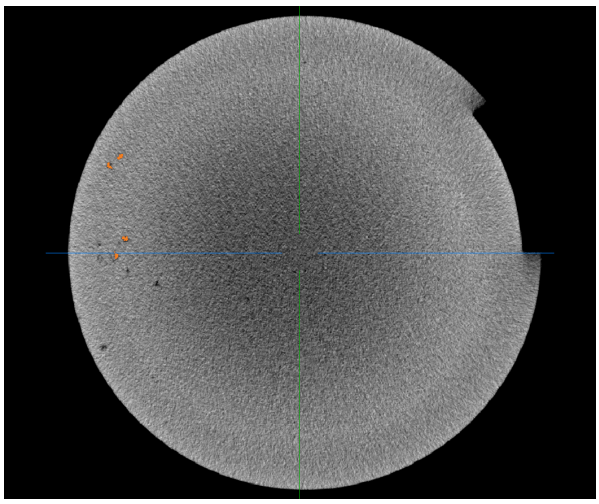
Due to this manual intervention, the described workflow is not suited to label the large amount of data required for the development of an online monitoring CNN. Furthermore, it cannot be guaranteed that the resulting label map does not contain false positive labels as the manual refinement is not performed by a qualified inspector and did not focus on individual anomalies. Nonetheless, the result can be used as a starting point to train CNNs to analyze the CT volumes. The training and results of which are presented in Section 4.5 and Section 5.1 respectively.

Label Map Optimization

In the second step, the developed CT CNN is used to analyze a variety of CT volumes. As shown in Section 5.1, the CNN is able to detect 226 out of 228 anomalies in sample A11, with only two false



(a) Cross section of a CT volume with pores visible on the left side of the sample.



(b) The CT cross-section, with the label map shown in orange.

Figure 4.13 Exemplary CT cross-section with the corresponding label map.

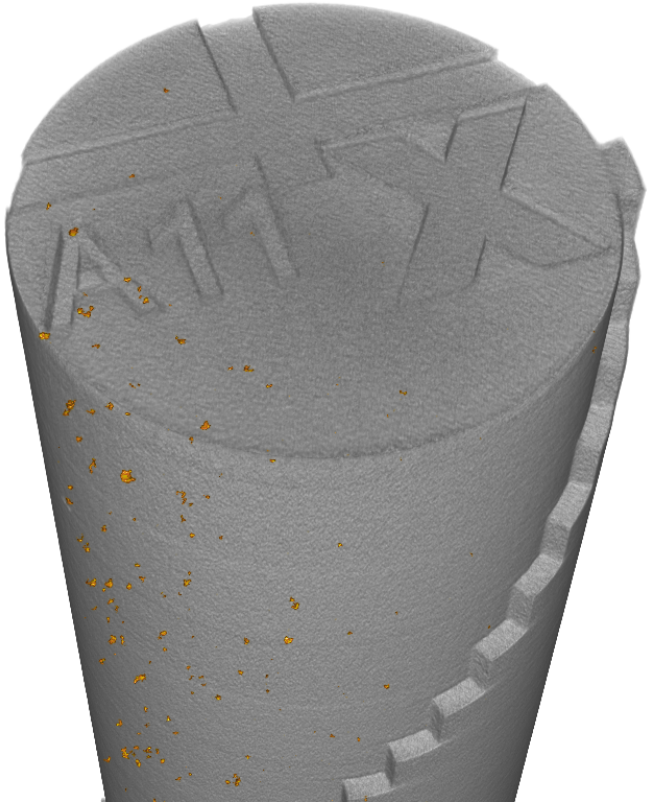


Figure 4.14 The CT volume visualized in 3D with the labeled defects shown in orange. The CNN is able to obtain the information in 3D and, therefore, is able to predict defects based on the volumetric data rather than single slices.

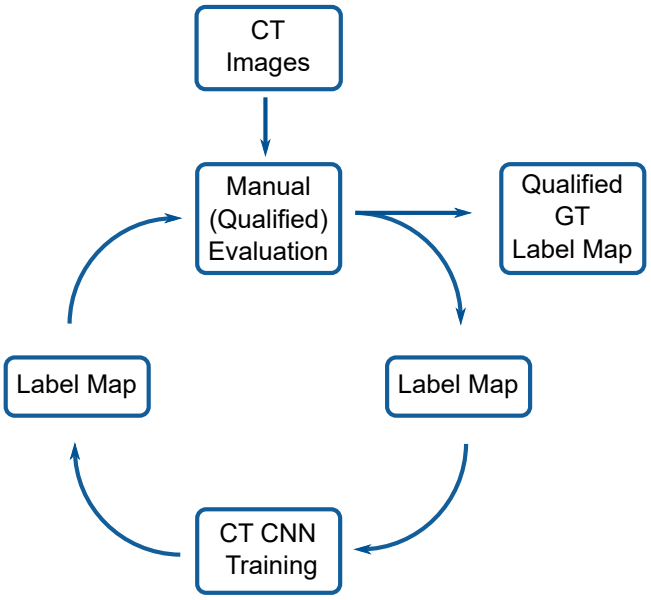


Figure 4.15 Visualization of the segmentation workflow and the iterations taken to produce a qualified GT label map together with a CNN for CT segmentation.

negatives but over 700 false positives. Hence, the CNN performed well in detecting anomalies but has to be improved with regard to the false positive rate. In a third step, the results of the intermediate CT CNN are used as input for a qualified optimization of the GT labels. For this optimization, qualified CT experts inspected the CT volume and the corresponding label map. They analyzed the volume on a per-slice basis and deleted falsely labeled (false positive) voxels or added falsely missed (false negative) voxels to the GT label map. Depending on the number of anomalies and the quality of the label map produced by the CT CNN, this took around twelve hours for 600 slices. The Step Samples used in this study consists of approximately 3300 slices. If a sample could not be labeled entirely due to time limitations, the CT volume and label map are subsequently cropped to the region analyzed by the qualified inspector.

4.3.4 Qualified Label Map

The label maps refined by the qualified CT inspectors are referred to as qualified label maps in the following. CT inspection is an inherently challenging task. The differentiation between a real defect and artifacts or irrelevant anomalies in low-contrast regions is not strictly rule-based but requires the experience of a trained human inspector. Particularly in aviation, the strict safety standards require a qualified inspector to take this decision. While the inspector follows a given standardized procedure for the CT analysis, this procedure cannot be translated to clear algorithmic rules for segmentation. If this were the case, the use of CNNs would not be necessary in the first place, as the analysis of CT volumes could be solely based on conventional rule-based algorithms. Instead, the current procedure focuses on the human labeling aspect and provides the inspector with guidelines on how to measure a defect on a macro level. This becomes apparent in the following examples.

Firstly, the task of sizing and segmenting a pore represents two different modalities. While the inspector focuses on sizing the entire defect on the macro level, the introduced CNN approach segments the entire volume on the voxel level. In particular, the inspector measures the pore using a virtual ruler, which does not require the labeling of each individual voxel. The CNN, on the other hand, classifies each individual voxel with regard to its defect or non-defect type. This provides a more refined analysis, which is beneficial but is not clearly defined in the current procedures. For example, there is no definitive rule for the border of a defect on the voxel level. The difference between the defined procedure for labeling a pore and the more detailed segmentation by the developed CNNs is visualized in Figure 4.17.

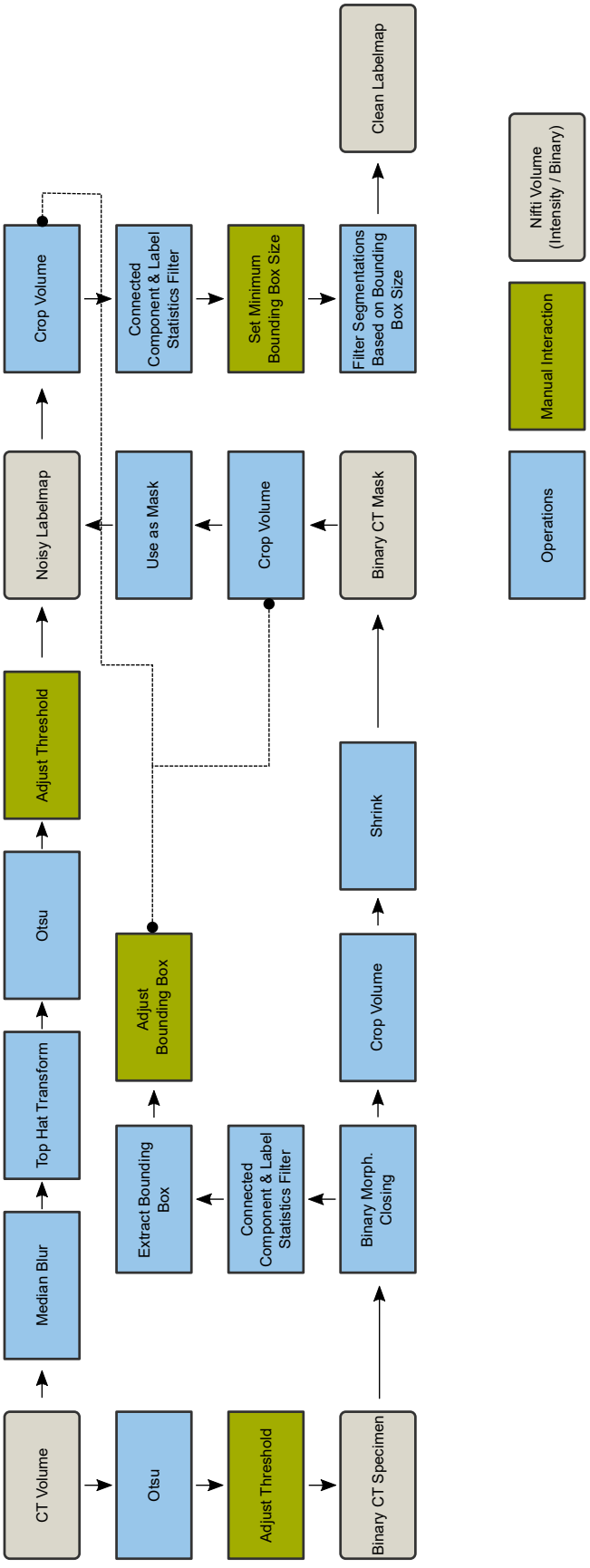
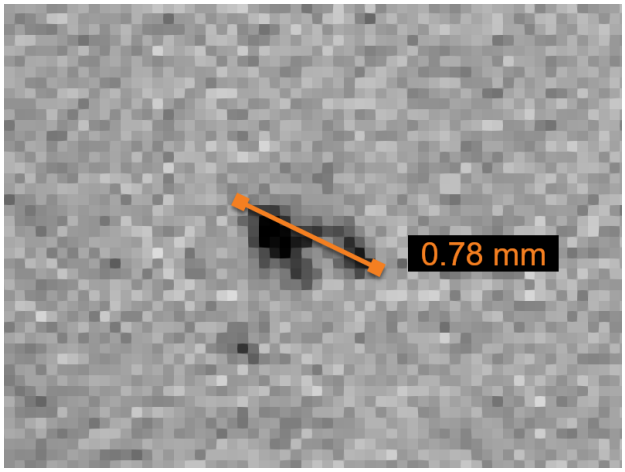
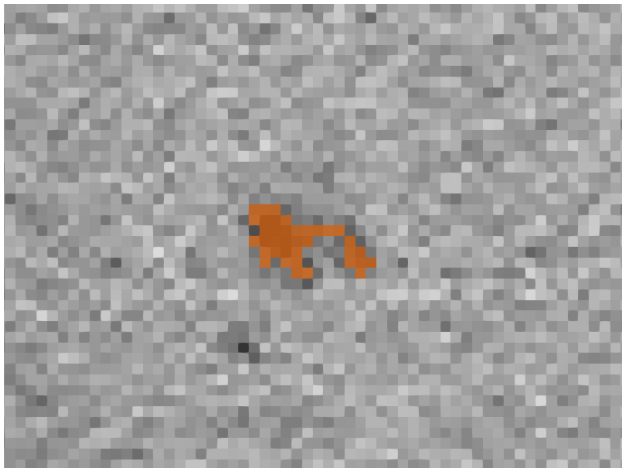


Figure 4.16 Workflow for the semi-automatic segmentation of anomalies in CT data (adapted from [Permadi, 2021]).



(a) A pore visible in the CT image. The ruler illustrates the way a qualified inspector might measure the defect (indicated here only for illustration purposes by non-expert).



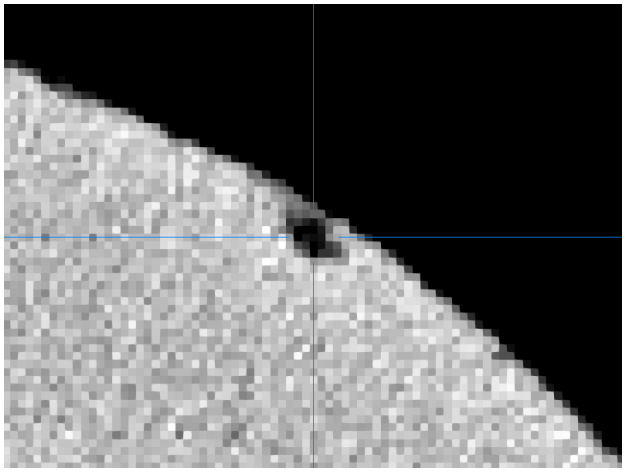
(b) The same pore as on the left but this time with the segmentation label map overlaid as it is used in this work.

Figure 4.17 Illustration of the difference between the qualified procedure and the more refined but also more complex segmentation label map used in this work.

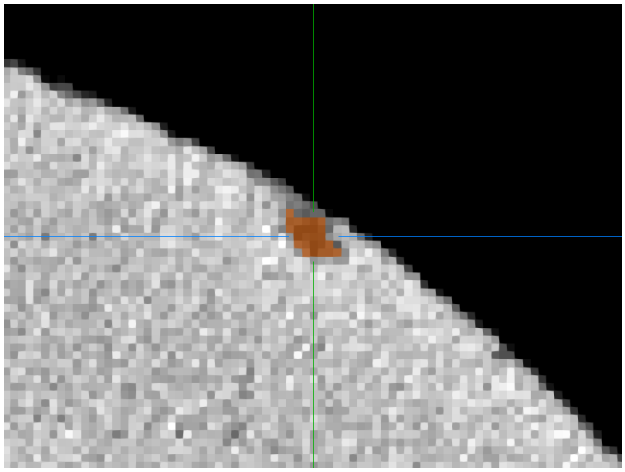
Secondly, besides the representation of a pore on a macro level, the current procedure provides the human inspector the possibility to mark regions that he deems as clear defects even if only one of the multiple indications for a defect is fulfilled. For example, even if the inspector is not required to mark a pore that is open to the surface (as such pores will be detected in a later penetrant testing step), he can label such a pore. This procedure is logical and beneficial for human inspection. On the other hand, it produces an additional challenge for an automatic algorithmic-based analysis as the line between an anomaly being labeled as defective or non-defective is not entirely defined. An example of such an anomaly is shown in Figure 4.18.

Thirdly, on a legal aspect, the inspector also has to take responsibility for the inspected part and might be liable in case of mistakes. Currently, in aviation, this responsibility cannot be transferred to an algorithm, in particular not to an ML algorithm, due to a lack of legal framework. The EASA is actively working on the required legal framework and aims to approve the first ML applications by 2025 [EASA, 2020]. As a result of these human-oriented inspection rules, the created segmentation label maps may contain ambiguous label cases. These ambiguous labels might impede the training of neural networks as the network is hindered in learning the exact rule by which to distinguish between defect and non-defect. Even more importantly, the evaluation of the network performance relies on the GT label map. As is explained in Section 4.4.3, the performance of the CNN is evaluated based on use-case-specific metrics that compare the prediction of the CNN to the GT. This comparison is inherently difficult for the border cases described above.

Nevertheless, while the current inspection rules cannot be applied one-to-one to the task of segmentation, the voxelwise labeling of the CT volumes by qualified inspectors represents the gold standard for CT analysis. The described ambiguous labels only make up a small fraction of the overall defect count, and the quality of the qualified label maps can still be regarded as high. In particular, the qualified label maps incorporate the expert knowledge of the qualified inspectors into the evaluation process. Combined with a qualitative visual inspection of the CNN results, a well-founded evaluation and comparison of the CNN performance is feasible under the described constraints. For the analysis in the following chapters, this visual inspection will be highlighted for selected cases to showcase some ambiguous cases. In addition, based on these findings, the current procedure for the CT evaluation will be revised in the future to better account for the automatic analysis by computer vision.



(a) CT image of a pore open to the surface of the specimen.



(b) The same pore as on the left with the segmentation label map overlaid.

Figure 4.18 Pores that are open to the surface do not have to be detected by CT as they are detected by a subsequent penetrant testing step. Nevertheless, they can be marked by an inspector in the CT if deemed relevant. This introduces an ambiguity to the labeled data.

4.4 Quantification Criteria

To train and evaluate neural networks, case-specific metrics have to be developed. The selection of the metric is critical for the training as well as the correct assessment of the network performance and, therefore, has to be justified to fit the given requirements [Virkkunen et al., 2021]. In the following, the selection and implementation of the metrics used in this thesis are presented, as well as their applicability justified.

4.4.1 Deep Learning Metrics

For the training of CNNs, there exist a variety of metrics. For the given task of image segmentation, one of the most prominent is the Dice score and variations of it [Reinke et al., 2021].

Dice Loss

The Dice score (or Dice-Sorensen-Score) uses the overlap of two volumes as the main criteria for evaluation. It was independently introduced by Dice [1945] and Sørensen [1948] and is computed by:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.1}$$

Where A and B are the two areas to be compared, and $|A \cap B|$ is the overlap of the two volumes. A Dice score of 100% corresponds to a perfect overlap of the two volumes. A Dice score of 0% means no overlap at all. Figure 4.19 visualizes the concept of the Dice score.

The Dice score can also be rewritten as a loss function for CNN training [Sudre et al., 2017; Milletari et al., 2016]:

$$DiceLoss(A, B) = 1 - \frac{2|A \cap B| + \epsilon}{|A| + |B| + \epsilon} \tag{4.2}$$

Where ϵ is a small constant added to both terms to avoid numerical issues by dividing by 0. Despite its known limitations the Dice score is a commonly used metric for CNN training [Reinke et al., 2021]. In particular, its quick and well-implemented computation facilitates its use in machine learning as it has to be computed hundreds of times during network optimization. Furthermore, it was shown to perform relatively well for segmentation tasks of small regions in large volumes [Sudre et al., 2017; Milletari et al., 2016].

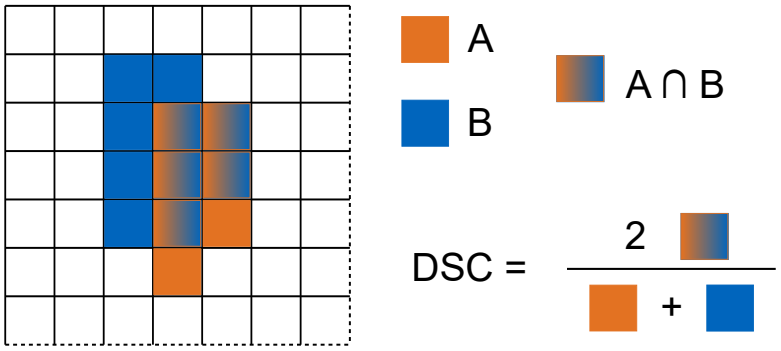


Figure 4.19 Visualization of the Dice score (adapted from [Reinke et al., 2021]).

Generalized Dice Loss

The Generalized Dice score is an extension of the Dice score first presented by Crum et al. [2006] and adopted as a loss function by Sudre et al. [2017]. The Generalized Dice Loss (GDL) is defined as:

$$GDL = 1 - 2\frac{\sum_{l=1}^{2} w_l \sum_n r_{ln} p_{ln}}{\sum_{l=1}^{2} w_l \sum_n r_{ln} + p_{ln}} \tag{4.3}$$

Where $r_l n$ is the voxel value of the reference volume, $p_l n$ is the prediction, and w_l is a weighing factor used to provide invariance to different label set properties. The weighing factor decides how much each class contributes to the GDL. In the case of $w_l = 1/(\sum_{n=1}^{N} r_l n)^2$, each class label is weighted by the inverse of its volume. This is supposed to reduce the correlation between region size and Dice score and hence increase model performance for highly imbalanced datasets [Sudre et al., 2017].

4.4.2 Probability of Detection

The Probability of Detection (POD) is a standard metric in the qualification of NDT systems. It is defined as the proportion of times the system will detect flaws of a given size [Hovey and Berens, 1988]. The POD of an ideal inspection system would be a step function depending on the flaw size. As the goal of the inspection is to detect all flaws greater than a specified threshold, this ideal setup would detect no flaws below the threshold and then step up to detect all flaws greater than the threshold. The size threshold can be defined by, e.g., material models or fatigue tests. For a realistic NDT setup, the POD cannot be assumed to be a step function, as there exists a variety of influencing factors that can lead to errors or imperfections in the test program. Additionally, only a limited amount of tests can be performed, limiting the exact assessment of the system's capabilities. Therefore, the POD is described statistically [Hovey and Berens, 1988]. It should be noted that there exist different methods and algorithms to obtain the POD of a system.

For this work, the statistical estimation is performed according to the conventional aerospace standards [Berens and Hovey, 1984; Hovey and Berens, 1988; Schnars and Kück, 2009]. Here, it is assumed that the response caused by a single anomaly correlates with its size, i.e., a larger anomaly should lead to a stronger response of the system. For the presented monitoring system, this is assumed to hold true as a larger anomaly is expected to influence the meltpool more strongly, which in turn should lead to a stronger signal deviation of the photodiode. The following analysis by deep learning is expected to be sensitive to this behavior, as it depends on the data obtained by the photodiodes. Theoretically, the performance of the trained CNN could be biased by the training data if the training dataset is not large and diverse enough to allow for sufficient generalization of the network. This could influence the sensitivity of the overall setup to different anomaly sizes. In particular, a CNN trained solely on small pores might be unable to detect large pores as it could not generalize the trained concept to larger pore sizes. This behavior is investigated empirically in Section 5.2.5 and cannot be observed for the study at hand. Therefore, the described dependency is assumed to hold true.

As a conservative approach for safety-critical applications, a confidence bound is introduced in the POD calculation. This bound shall account for random errors in the test program and adds an additional safety margin to the calculated POD curve. The aerospace industry usually uses the 95 % lower confidence bound [Hovey and Berens, 1988; Schnars and Kück, 2009]. This implies that if the POD would be calculated for every sample ever produced in 95% of the cases, the POD would be as high or higher as the 95% confidence bound [Härdle et al., 2015; Cheng and Iles, 1988].

Figure 4.20 shows an exemplary POD curve and the 95% confidence bound. The relevant metric for the evaluation of the inspection system is the x-value of the intersection of the confidence bound with the 90% on the y-axis. It is referred to as the $a_{90/95}$ (in the following for simplicity reasons only POD) and represents the minimum size of a defect that the system would still detect with a POD of 90% with a confidence of 95%.

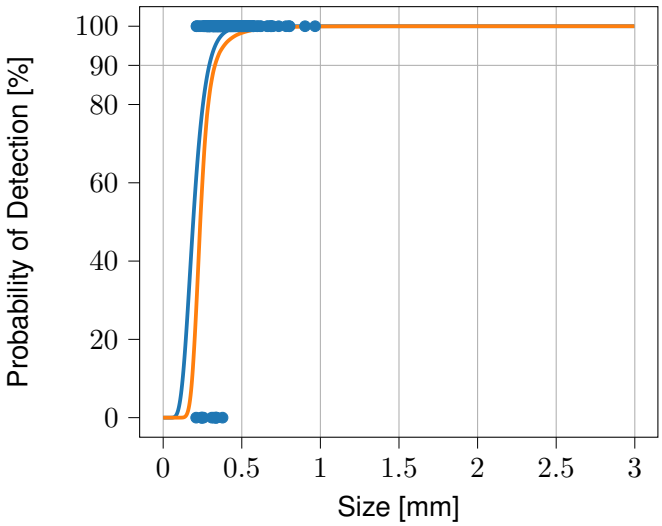


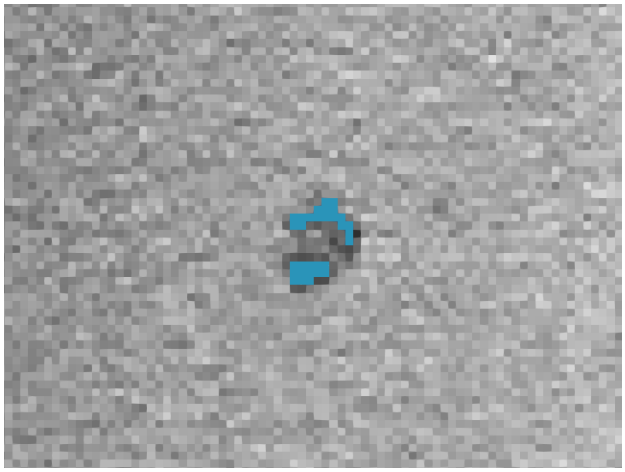
Figure 4.20 Exemplary POD curve (blue) and the 95% confidence bound (orange). The POD ($a_{90/95}$) as a relevant metric for the evaluation of the inspection system is the x-value of the intersection of the 95% confidence bound with 90% on the y-axis (here approximately 0.4 mm).

4.4.3 Hit-Miss Analysis

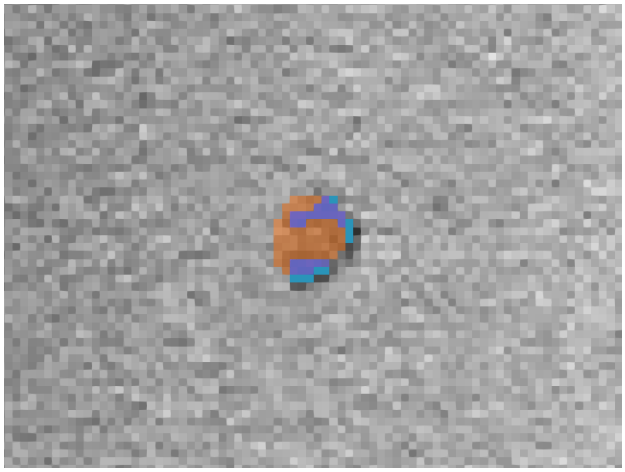
As described above, the POD is calculated based on the size of the detected and not detected anomalies. Hence, this information has to be extracted from the segmentation label map. For this, each ground truth anomaly is measured by drawing a bounding box around it. The bounding box is oriented along the three major axis of the anomaly. The size of the anomaly is then extracted by measuring the largest dimension of the bounding box.

To determine whether the anomaly is detected by the model, a comparison of the ground truth label map and the prediction label map is performed. In the first step, the predicted labels are dilated by three voxels in the xy-direction and six voxels in the z-direction. In the second step, each ground truth label is checked and is marked as detected if one voxel of the dilated predicted label map intersects with one voxel of the ground truth. With a voxel size of 60 µm, this results in an increased area of influence per prediction of approximately 180 µm in each direction in the xy-plane and approximately 360 µm in the z-direction. The increased area of influence is supposed to take inaccuracies in the pre- and post-processing into account. In particular, the registration of the CT volume to the OM volume is a known factor to introduce geometrical inaccuracy. Here, the average error of the registration is estimated to be around 380 µm [Montijo Badeira, 2020].

Besides the known inaccuracies, the dilation is also motivated by the fact that an exact positioning of the anomaly is not critical to the application. In contrast, for the use case of image segmentation for autonomous driving, the position of an object (e.g., a pedestrian) is highly critical. For the detection of an



(a) CT image with two GT labels (blue) in close proximity.



(b) The same CT image with the two GT labels (blue) and an exemplary prediction of a CNN (orange) that spans both GT labels. The overlap of both label maps is visualized in purple.

Figure 4.21 The hit-miss analysis does not perform a one-to-one mapping but marks all GT labels in the proximity of the prediction as found. In the illustrated case, both GT labels are declared as detected.

anomaly in the AM sample, on the other hand, the detection and sizing of the anomaly are more relevant than its precise localization. Hence, the increased influence sphere is chosen as a way to account for the inaccuracies and the given use case.

It is important to note that this hit-miss analysis does not perform a one-to-one mapping of labels. In particular, it does not match individual prediction labels with specific GT labels. Instead, it is possible that multiple GT labels are marked as found by one inference label if the proximity criteria described above are fulfilled for each of them. This is illustrated in Figure 4.21. Both defects in the GT label map are marked as found even though only one inference label lies in their proximity. On the one hand, this may lead to the unintentional marking of small defects as found if they occur independently in the proximity of a larger defect that is actually detected. On the other hand, the same behavior is desirable when multiple small defects occur as a result of one larger process deviation. In this case, the one process deviation detected by the CNN corresponds to the accumulation of smaller pores. As the formation of individual pores in the investigated specimens is unknown, no one-to-one mapping is performed.

This use-case-specific quantification adds valuable information to the standard validation and testing procedures in deep learning, which mostly operate on common metrics such as the Dice coefficient or intersection-over-union. As pointed out in Section 4.4.2, the hit-miss analysis and POD calculation are standard methods for the qualification of NDT methods. Transferring this well-established statistical method to the quantification of CNNs is a first step towards a possible qualification of CNNs in NDT. However, it should be stressed again that the POD requires a correlation between the anomaly size and its detection probability (see Section 4.4.2). The empirical investigations support this assumption for the investigated use case, but it is not proven formally as described in Section 4.4.2.

Besides the POD, the "Biggest Undetected Defect" (BUD)/ "Biggest False Negative" (BFN), the "Number of Undetected Defects" (NUD)/ "Number of False Negative" (FN), the "Biggest False Positive" (BFP) and the "Number of False Positives" (FP) are defined as relevant metrics for the quantification of the network performance. Additionally, the "Number of Undetected Defects larger than 400 µm" (NUD_{400}) and the "Number of False Positives larger than 400 µm" (FP_{400}) describe the respective metric for the relevant defect sizes. For the comparison of results on different specimens, the NUD_{400} is also expressed as the ratio of undetected defects larger 400 µm to all defects larger than 400 µm. All metrics can be calculated on the same data basis as the POD, but in contrast to the POD, those metrics can be extracted always. Hence, they provide valuable additional information about the network performance. In particular, the FP_{400} represents an important supplement to the POD as the size and number of false positives are not regarded

in the pure POD calculation. Nevertheless, a system with a perfect POD but a very high number of false positives is of limited practical use as the high number of ill-founded rejections prohibits its application in reality. Additionally, the BUD and NUD contribute valuable information in the case when no POD can be calculated or when the statistical distribution of the detected/ undetected anomalies influences the logistic regression in the POD calculation significantly.

The calculation of the introduced metrics is computationally expensive and not optimized to be integrated into the training process itself. Therefore, the calculation is performed in parallel to the validation on the epoch level. In this work, it will be referred to as pseudo-testing. "Testing" as it tests the model in reference to the later relevant metric. "Pseudo" as it cannot be regarded as complete testing because it is run in parallel per epoch and hence influences the decision about which model and checkpoint is chosen. For thorough testing of the model performance, the testing data should be different from all previous datasets influencing the training or selection of the model.

4.5 Training Process Computed Tomography

In the scope of this thesis, a large variety of neural networks is trained and tested. As it would exceed the scope of this thesis, only a selection of the most relevant ones with meaningful results is listed here.

The CT CNNs fulfill a double role in this study. On the one hand, they are an essential prerequisite for the training and analysis of the OM networks as they are used to create the GT label map for the OM CNN training. On the other hand, they constitute their own field of research with strong links to the medical imaging sector as described in Section 3.1. Hence, the training and results presented in the following describe their own separate research contribution and are therefore split into the Methods (Section 4.5) and Result (Section 5.1) chapter as well. For enhanced readability, selected results might be anticipated to allow for a logical presentation of the iterative procedure detailed in Figure 4.15. The results are summarized in Section 5.1.4.

4.5.1 1. CT U-Net: Baseline

The following details the CT CNN for the first iteration. This is not the first CNN trained in the scope of this thesis, but it constitutes the starting point for the iterative CT label map optimization. For reasons of readability, previous CNN trainings are not outlined in detail. Instead, the inclined reader is referred to [Holtmann et al., 2021a,b; Permadi, 2021]. Due to the higher complexity of pore segmentation compared to the segmentation of inclusions, the first iteration focused solely on pores as anomalies.

Architecture As discussed in Section 2.3 and Section 3.1, the U-Net architecture and adaptations of it represent state-of-the-art neural network architectures for image segmentation, i.e., 3D volume segmentation in medical imaging. Hence, the general setup and architectures are transferred to CT segmentation of pores and inclusions. The U-Net expects a 3-dimensional input (i.e., a nifti volume) with only one channel (CT data contains only one information channel). It comprises five contracting blocks and five expanding blocks with 16, 32, 64, 128, and 256 channels, respectively (see Figure 4.22). Skip connections are implemented between the contracting and corresponding expanding blocks. The CNN is trained solely on pores, so the network has only two output channels (pore vs. background).

Training Data The dataset used for training consists of five samples of the Buildjob 500, which are printed with similar parameters as Buildjob B (reduced focus). It is only used for the training of the CT CNN, as the online monitoring system was not running correctly during the print job. For more details, see Appendix 2. The label maps used for the training are annotated as described in Section 4.3.3.

Based on these label maps, the five volumes are divided into smaller patches for training. To balance the two classes anomaly vs. no-anomaly, the patches are cropped such that eight out of ten patches contain an anomaly in the center of the patch. This facilitates the training process by a more efficient data loading

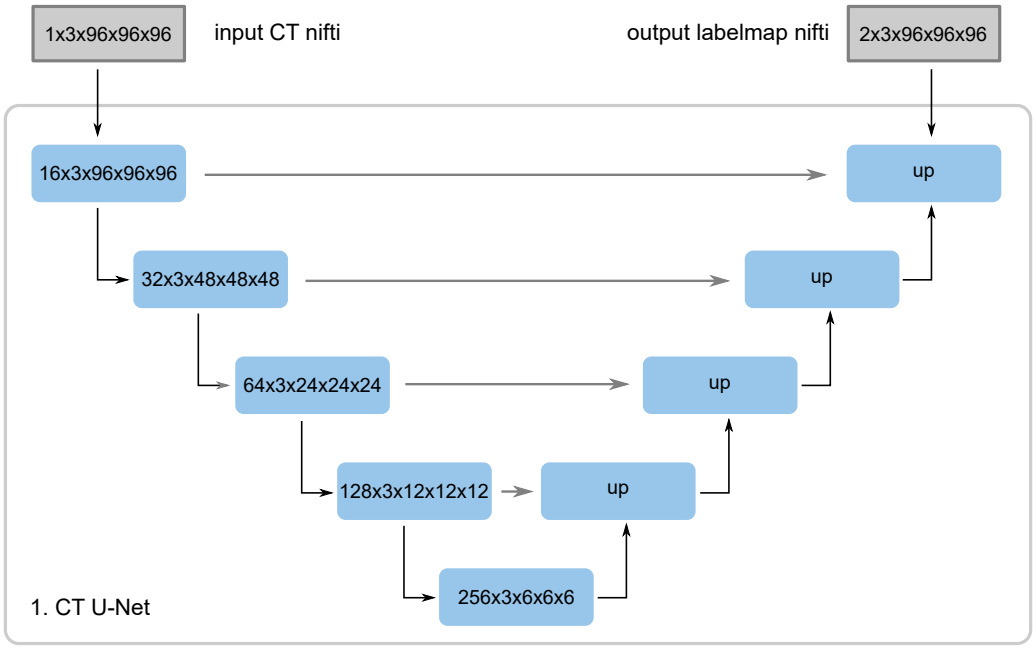


Figure 4.22 Visualization of the U-Net architecture for the 1. CT CNN.

process as well as focusing on the anomalous regions. The number of voxels containing an anomaly in the entire CT volume is significantly smaller than those without an anomaly. In a classification task, this is called class imbalance. The cropping results in a more balanced dataset even though the volume ratio of non-anomalous regions is still dominant. The cropped patches have a size of 96 x 96 x 96 voxels. A total of 2044 patches are cropped per sample. The center is chosen randomly with the boundary condition of an 8:2 mix, as described above. This means that one anomaly or region could be cropped multiple times or, by chance, never.

To maintain the global image properties for all patches and to improve training performance, the intensity of the CT image is normalized by subtracting the global mean of the image intensity and then dividing it by the global standard deviation (σ). Normalizing the input data is best practice for training CNNs as it facilitates the learning process [LeCun et al., 1998]. As all volumes contain regions of air, a common reference value for the CT intensity is known. This ensures the transferability between different samples.

As an additional pre-processing step, the patches are altered by a random affine transform. With a probability of $50\,\%$, a patch is translated by up to $\pm 10\,\%$ in each spatial direction. Additionally, it is rotated around each axis by up to $90\,°$ and scaled by a factor of up to $\pm 10\,\%$ with the same probability. This data augmentation aims to improve the generalization of the CNN as the variety of anomaly sizes and shapes is artificially enlarged. This also reduces the chances of the network overfitting to the training dataset [Wang et al., 2020; Wong et al., 2016; Shorten and Khoshgoftaar, 2019].

Validation Data Two samples from different buildjobs are used to validate the network during training. One sample (SmallSteps 06) is produced with the parameter set described in Section 4.1 Buildjob A and the other one (Steps 313) belongs to Buildjob 300 described in Appendix 2. The label map for the sample SmallSteps 06 is created from scratch by manual labeling by a not-qualified inspector. The sample Steps 313 is analyzed by the semi-automatic labeling workflow and subsequently checked and optimized manually. The validation dataset consists of 1533 patches, which are created by the identical cropping mechanism used for the training dataset. In contrast to the training data, the validation patches are not augmented by a random affine transform.

Training Configuration The network is trained with Adam optimizer with a learning rate of 10^{-4} to minimize the Dice loss [Kingma and Ba, 2017]. For this, the U-Net is fed with batches of 32 patches.

The intermediate feature maps are batch normalized. Validation is performed every epoch by calculating the mean Dice over all the validation patches. For this purpose, inference is performed on all validation patches with a batch size of four. Training and evaluation are implemented in monai and Pytorch Lightning with a maximum of 50 epochs [William Falcon et al., 2020]. The results are shown in Section 5.1.1.

4.5.2 2. CT U-Net: Pores & Inclusions

The second CT CNN to be introduced incorporates inclusions in the dataset. Additionally, it uses the findings of the 1. CT CNN to improve the model performance, i.e., regarding the high false positive rate and the influence of CT artifacts. For a more fine-grained assessment of the iteration process, the inclined reader is referred to Permadi [2021]. The model is developed prior to the qualified GT labeling and, therefore, also uses the semi-automated labeling workflow. The presented CNN is trained to detect pores and inclusions as separate anomalies.

Architecture The general U-Net architecture is adopted from Section 4.5.1 with the following adjustments. The model takes a 3D volume with one channel as an input and returns a three-channel output (pore vs. inclusion vs. background) of the same dimension. It comprises five contracting/ expanding blocks of channel sizes 16, 32, 64, 128, and 256. The number of residual units is set to two.

Training Data The training dataset consists of five samples; four samples (Sample 509, Sample 510, Sample 511, Sample 512) contain only pores, and one sample (X27) contains only inclusions as anomalies. Using four pore specimens and only one inclusion specimen is supposed to counteract the class imbalance and difference in training complexity otherwise introduced by the number of anomalies per sample. As pointed out in Section 4.5.1, the two datasets can only be used to train the CT models as the monitoring system did not record the meltpool data properly. The buildjobs properties are presented in Appendix 2.

The most relevant changes introduced by the 2. CT CNN can be attributed to enhanced data labeling. While the 1. CT CNN focused on a broad and diverse training dataset, the 2. CT CNN reduces the amount of training data in favor of a higher-quality of the label map. To achieve this, an improved labeling workflow is introduced tailored to tackle the previously described challenges arising from CT artifacts.

Firstly, the volume is intensity normalized and then scaled to the range between 0 and 1. Subsequently, the volume is manually cropped to the middle section (in z-direction) of the specimen. The semi-automatic labeling process described in Section 4.3.3 shows considerably better segmentation performance for this region. Hence, this step reduces the number of false positives in the GT label map, often located at the top and bottom of the sample.

Secondly, all anomalies smaller than $0.2\,\mathrm{mm}$ are removed by an image filter. Additionally, false positives around the steps surrounding the sample (as shown in Figure 5.8) are erased manually.

Thirdly, the inner radius of the specimen is excluded from the segmentation. As described in Section 5.1.1, the inner part of the sample is prone to CT artifacts due to beam hardening. Hence, the semi-automatic algorithm has difficulties labeling these regions precisely and is more likely to create false positive labels. By excluding the center of the specimen, those false positives are excluded from the training data, reducing the likelihood of the model adapting to those labels. To prevent the CNN from misinterpreting the center of the specimen to always be anomaly-free, no training patches are sampled from this region. Instead, patches are cropped to always contain an anomaly, i.e., the entire sample is cropped such that the middle of a patch is defined by the middle of an anomaly. Therefore, the training dataset consists only of patches with an anomaly in the center. Patches may contain additional anomalies in the proximity of the center anomaly but never no anomaly.

The steps described up to now focused on preventing false positives by limiting the region of interest for labeling to the parts of the specimen in which the semi-automatic workflow shows high accuracy. By doing so, the quality of the label map is improved while reducing the number and variance of anomalies. To counteract the limitation in shape and geometric variance in the training data, an extensive data augmentation

is introduced in the fourth step. I.e., the patches are augmented by a random contrast adjustment and a random affine and elastic transformation. For an in-depth analysis of the chosen augmentation parameters and the empirical results, see Permadi [2021] and Simard et al. [2003]. In total, 4215 training patches of size 96x96x96 voxels are generated. Of these 4215 patches, 4088 contain a pore, and 127 patches are cropped from the inclusion sample. The high imbalance in patches is due to the number of anomalies per patch and the varying complexity imposed by the two different anomaly types. While pore patches contain only a few but complex anomalies, the inclusion patches may contain many anomalies of mostly similar size and shape. Hence, it is easier for the CNN to adapt to inclusions, as is shown in Section 5.1.2.

Validation Data The validation dataset consists of four samples: three specimens containing pores (SmallSteps 06, A10, and A22) that are manually optimized and one specimen with inclusions (X27) labeled semi-automatic. Therefore, the validation dataset combines different anomaly types from different samples. The volumes are cropped to 96x96x96 voxel patches for better performance of the training and validation routine. The validation patches are not augmented prior to evaluation. The validation dataset consists of 851 patches (510 pores and 341 inclusion), which is around 20% of the training dataset.

Training Configuration The training is split into two parts to optimize the adaption of the model to both anomaly types. For both stages, the Adam optimizer with a learning rate of 0.001 is used. The objective is to minimize the mean Dice loss for the two anomaly classes. In addition, the mean Dice loss for the validation data and the validation Dice scores for the two classes are tracked individually to analyze the training performance. In the first training phase, up to Epoch 40, the training focused on the more complex pore anomalies by strongly oversampling pore patches over inclusions. This is achieved by the training data setup as described above. In the second phase, from Epoch 40 to 60, the classes are sampled more evenly to allow for better training of inclusions. The patches are sampled randomly from both classes and combined into batches of sizes 32 and 4 for training and validation. The training process is implemented in Pytorch Lightning and monai, with validation performed every epoch.

4.5.3 3. CT U-Net: Performance Optimization

To further improve the CT CNN, a selection of the qualified label maps is used to finetune the 2. CT CNN. For this, the network architecture and results of the 2. CT CNN are used as the basis for transfer learning. By incorporating parts of the qualified data the CNN is pushed closer to the "qualified" interpretation of defects as labeled by the inspectors rather than the unqualified interpretation by students and non-experts. At the same time, the 3. CT CNN aims to retain as much information as possible from the previous experiments by using a transfer learning approach. This should allow for a quick adaptation of the network as it is already trained on the data described in Section 4.5.2.

Architecture The network setup is equivalent to the 2. CT CNN (Section 4.5.2) as it was already optimized in Permadi [2021]. This allows for an easy adoption of the previous model weights by transfer learning.

Training Data As the model structure is not altered between the 2. and 3. CT CNN, the expected performance improvement has to be achieved by optimizing the training data. In order to further reduce the number of false positives while maintaining a low POD and false negative score, parts of the qualified CT data are incorporated into the training. The samples B17, D05, D06, D14 and D18 are used for training. All specimens only contain pores and no inclusions.

To train the model more efficiently, only challenging areas of the samples are included in the training. The patches are selected on the basis of the 2. CT CNN results. In particular, all areas where the 2. CT CNN detected an anomaly are chosen as training samples by cropping patches of 120x120x120 voxels around them. The corresponding label maps are cropped from the qualified GT label map with the same

origins and dimensions. Hence, the resulting patches are defined by the previous model's segmentation but contain the qualified dataset's GT information. Therefore, if the model predicts an anomaly correctly, it will be reinforced by being shown more actual anomalies from the new dataset. On the other hand, if the prediction is a false positive, the CNN can adapt to this incorrect segmentation as this region is shown with the correct label map from the GT.

This setup resulted in 4200 training patches, which are pre-processed by intensity normalization based on the entire volume. Subsequently, smaller cubes of size 96x96x96 voxels are randomly cropped from the intermediate cubes. This allows for a more diverse orientation of the anomaly within the training cube. Without randomly cropping the volume a second time, the anomaly would always be centered in the volume. This might influence the training process as the network could adapt to this fact and always predict anomalies in the center of a volume. Additionally, the patches are augmented by random contrast adjustment and random elastic deformation.

Validation Data The samples B23 and D26 are used for validation. The validation patches are generated the same way as the training patches. This resulted in 2776 validation patches, which are intensity normalized but not augmented further.

Training Configuration The general training setup is adopted from Section 4.5.2, but the training is conducted in one phase as no additional data of inclusions is used. The objective of the Adam optimizer is the minimization of the mean Dice loss. In contrast to the previous training, the learning rate is decreased to 0.00005 to allow for more refined transfer learning. A small learning rate should facilitate the preservation of the previously gained information of the CNN by slightly reducing the impact of new training data. The assumption is that the training of the 2. CT CNN already allowed the network to adapt to the general concept of anomalies. Using this pre-trained network and only finetuning the weights to the new data should lead to a better-performing model with less training effort.

4.6 Training Process Online Monitoring

Based on the results gained by the CT CNNs described above and in Section 5.1, in the following, the design and training of CNNs for detecting anomalies based on the online monitoring data are presented. As for the CT CNNs, the network architecture, training setup, and data pre-processing are described in Section 4.6, and the results are shown in Section 5.2. Different CNN trainings are investigated and compared based on the metrics defined in Section 4.4. Listing all investigated networks and training setups would exceed the scope of this work. Hence, only a selection of well-performing or meaningful CNNs is given. The 1. OM U-Net and 2. OM U-Net show the general feasibility of the proposed method to evaluate OM data by CNNs. Here, an in-depth analysis of the CNN performance and the influence of data augmentation on the training process is conducted. The 3. OM U-Net focuses on the influence of individual sensors and channels. It aims to provide a better understanding of the required input data and its influence on the prediction performance. The 4. OM U-Net investigates the transferability of the trained network from one defect creation mode to another, while the 5. OM U-Net focuses on the transferability between different defect sizes. In combination, the presented experiments aim to provide an in-depth understanding of the feasibility, robustness, and explainability of the developed CNN approach. This will be summarized and discussed in Section 6.2.

The nifti volumes described in Section 4.3.1 represent the data baseline for all following trainings if not specified separately. The ground truth information about the location and size of an anomaly is generated using the CT CNN presented in Section 5.1.3. Additionally, the qualified manual label maps are used for validation and testing.

4.6.1 1. OM U-Net: Baseline

In the first approach, the U-Net is trained on the samples of the buildjobs A and B. It is supposed to show the general feasibility of the proposed method and to deliver a first baseline for further comparison and optimization. The following describes the network architecture, the training data, including data augmentation, the validation data, and the training configuration (i.e., hyperparameter). This first approach follows a similar training pattern to the CT CNN training presented above.

Architecture The general U-Net architecture is adopted from the CT CNN trainings in Section 4.5, but in contrast to the CT CNN, the OM U-Net has a multi-channel input. While the CT image consists of only one channel (the gray value image of the CT scan), the OM domain can contain multiple channels (one for each sensor and pre-processing method). More details will be given in the paragraph "Training Data". In this case, six channels of the OM domain are used as input. The U-Net comprises five contracting blocks and five expanding blocks with 16, 32, 64, 128, and 256 channels. Batch normalization is used on the feature map level. The network returns a two-channel output for anomaly vs. background.

Training Data As described in Section 4.3.1 there exist different combinations of sensor and mapping methods. For the first approach, six different channels of online monitoring data are chosen for training. They consist of the three sensor readings (TED, TEP_{low}, TEP_{high}) with two track mappings each (minimum and maximum mapping). This results in the following channels in this order: TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, TEP_{low}^{min}. Hence, the information about the minimum and maximum at each voxel position for the three sensors is shown to the network. All channels are intensity scaled to a range of 0 to 1 independently from each other.

The ground truth label map is produced using the 3. CT CNN as described in Section 5.1.4. Using the ground truth, small cubes of size 120x120x120 voxels are center-cropped for each anomaly in the label map. Similar to the approach for the CT CNNs, this should facilitate the training process by balancing anomaly/ non-anomaly regions and speeding up data loading. To further increase training speed and push the network towards relevant defect sizes in this first approach, only cubes with anomalies larger than 0.5 mm are considered. Here, the size corresponds to the largest bounding box dimension. This results in around 2815 individual niftis.

Before being fed into the network for training, the cubes are augmented to enhance model robustness and prevent overfitting. Firstly, smaller cubes of size 96x96x96 voxels are randomly cropped from the intermediate cubes. Secondly, the intensity of the cropped cube is randomly shifted by $\pm 10\%$ to allow for more robust training. Thirdly, a random Gaussian noise (mean=0, std=0.05) is added to the volume for the same reason. The specimens and processing steps used for training are summarized in Table 4.3.

Validation Data For the creation of the validation data, the same 120x120x120 voxel cubes are created but from different specimens (A06, A12, A13, A30 and B06, B13, B28, B30). 2259 cubes are cropped from the eight validation specimens. This allows for a validation of the network performance on unknown data. To prevent the constant centering of the anomaly in the middle of the cube, smaller cubes of 96x96x96 voxels are cropped analogously to the training data. The specimens selected for validation are chosen from the different buildjobs such that each processing scheme is represented in the validation data, e.g., A12 for reduced laser power, B06 for focus shift, and B30 for skywriting delay. The cubes are used to calculate the validation Dice score.

In contrast to the cubes for the validation via Dice calculation, the samples A11, B17, and B23 are used for pseudo-testing. As described in Section 4.4, the relevant metrics for quantifying the model performance are the POD, BUD, NUD and FP. These metrics are unsuitable for a quick on-the-fly calculation and are, therefore, not integrated into the training process itself. Nevertheless, it is desirable to quantify the model performance per epoch for them. Hence, they are calculated in a post-process step and logged for each epoch. To achieve this, the prediction labelmap is generated and compared to the qualified label map as described in Section 4.4.3.

Table 4.3 Training data and pre-processing steps for the 1. OM CNN approach.

Training Data	
Channels	TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, TEP_{low}^{min}
Samples	A: 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 18, 19, 21, 22, 25, 26, 27 B: 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29
Number of Anomalies	2815
Data Pre-Processing/ Augmentation	Intensity Scaling (0–1)
	120x120x120 voxel centered cubes
	Sampling of anomalies larger 0.5 mm
	96x96x96 voxel random cubes
	Random Intensity Shift (offset: -0.1–0.1; probability: 0.5)
	Random Gaussian Noise (mean: 0; σ: 0.05; probability: 0.1)

Training Configuration The training process aims to reduce the Dice loss of the predicted probability map produced by a softmax activation on the raw model output. For this, the Adam optimizer with a fixed learning rate of 10^{-4} is used. The network is trained with a batch size of four. Batch normalization is performed during the training, which normalizes the intermediate results of the U-Net layers per batch.

Validation and pseudo-testing are performed every epoch with a validation batch size of four. The model is trained for 28 epochs, after which the training is terminated. As for the other networks, the training and validation are implemented in monai (Version 0.9.1) and Pytorch Lightning. The results are presented in Section 5.2.1.

4.6.2 2. OM U-Net: Performance Optimization

The second iteration closely follows the approach described in Section 4.6.1. It focuses on improving the model performance by introducing additional data augmentation. As discussed in Section 3.1, research suggests that the model performance depends more strongly on suitable data and training details than on the network architecture. Hence, the introduced improvements focus on finetuning the training process rather than finetuning well-established network architectures.

Architecture The network architecture is kept constant from the previous approach (see Section 4.6.1). This allows to better compare and evaluate the impact of increased data augmentation.

Training Data The basic data pre-processing is analogous to Section 4.6.1. All six channels (TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, TEP_{low}^{min}) are used as input to the network. The ground truth labels and the cropping of 120x120x120 cubes are the same as for Section 4.6.1, but in this case, all anomalies larger than 0.4 mm are considered for training. This results in 5661 individual cubes with an anomaly in the center. In addition to the previous training approach, the samples A20, A24, A28, and A29 are included in the training dataset, resulting in 627 additional training cubes.

As pointed out above, the main difference to the 1. OM U-Net lies in an increased data augmentation. Starting from the 120x120x120 voxel cubes, smaller cubes of size 96x96x96 voxels are cropped randomly. The resulting volumes are rotated and scaled randomly. Subsequently, the intensity and contrast are shifted randomly, and Gaussian noise is added to the volume. In addition to this conventional data

augmentation, the ground truth labels are dilated twice by one voxel in each direction. This enlarges the labels in each direction, increasing the volume/ voxel count labeled as an anomaly. The underlying idea for this augmentation is the high-class imbalance between anomaly and background. By increasing the volume of the defect label, an increased emphasis should be placed on anomaly regions, i.e., on originally small anomalies. The downside of this approach is the reduced geometrical accuracy of the label. The original label represents the actual structure of the anomaly as detected in the CT scan. When enlarging the label, this fine structural information is reduced or lost. This limitation is acceptable for the given use case of defect detection, as the exact geometric representation of the defect is of secondary importance. The sizing of the defect should still be possible within an acceptable variance as the enlargement induced by the data augmentation is known and can be subtracted in a post-processing step. The training data is summarized in Table 4.4.

Table 4.4 Training data and pre-processing steps for the 2. OM CNN approach.

Training Data	
Channels	$TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, TEP_{low}^{min}$
Samples	A: 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29 B: 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29
Number of Anomalies	5661
Data Pre-Processing/ Augmentation	Intensity Scaling (0–1)
	120x120x120 voxel centered cubes
	Sampling of anomalies larger than 0.4 mm
	96x96x96 voxel random cubes
	Random Affine (rotate: 0, 0, -180–180; scale: -0.2–0.2; probability: 1)
	Random Intensity Shift (offset: -0.2–0.2; probability: 1)
	Random Contrast Adjustment (gamma: 0.9–1.1; probability: 1)
	Random Gaussian Noise (mean: 0; σ: 0.05; probability: 0.5)
	Dilate Anomalies (size: 2 voxels; connectivity: 1)

Validation Data The validation data consists of the same cubes (A06, A12, A13, A30 and B06, B13, B28, B30) as introduced in Section 4.6.1. In addition to the cropping of 96x96x96 cubes, the labels are dilated analogously to the training data. This allows for calculating the Dice loss and coefficient on the same data basis. If the training labels were enlarged, but the validation labels were kept constant, the two label maps would be based on two different label assumptions, influencing the Dice negatively.

In contrast to the validation, the pseudo-testing on the qualified samples A11, B17, and B23 is not altered by increasing the labels. Firstly, this is not necessary, as the custom-designed metrics (BUD, NUD, FP) do not operate on a voxel level and, hence, are not influenced by the dilation of only one label map. Secondly, by not dilating the ground truth labels for pseudo-testing, the correct sizing of defects can be extracted directly from it.

Training Configuration The hyperparameters for the training are adapted from the 1. OM U-Net (see Section 4.6.1). By keeping the training parameters and network architecture constant, the influence of the increased data augmentation and database is investigated. The results are provided in Section 5.2.2.

4.6.3 3. OM U-Net: Individual Channels

The third approach aims to investigate the importance of the individual channels for the performance of the CNN. Per training only one channel is chosen as input for the model to get a better insight into the CNN and better understand the information incorporated in the different channels of the input data. To compare the individual channels, seven models with identical architectures are trained, one per channel plus one, taking all six channels as input. Models trained on a channel that does not contain relevant information should perform significantly worse than models trained on a meaningful channel. The degree of information per channel is estimated by comparing the individual CNN performances. This should allow for the selection of more relevant channels and sensors, which in the second step provides valuable information about the required sensor hardware and optimal pre-processing steps. This also facilitates the transfer to other printer systems with different sensor setups.

Architecture The previous U-Net architecture with five expanding blocks with 16, 32, 64, 128, and 256 channels respectively is used. It takes a one-channel input and applies batch normalization. It returns a two-channel output for anomaly or background.

Training Data The training data consists of specimens from Buildjob A. The specific specimens used for training are listed in Table 4.5. As for the first approach, in the first step, the data is scaled to a range between 0 and 1 and subsequently cropped into smaller cubes of 120x120x120 voxels. Each cube is cropped to contain an anomaly larger than $0.5\,\mathrm{mm}$ in its center. From those cubes, smaller cubes of side length 96 voxels are extracted. In the following data augmentation phase, the cubes are randomly deformed by an affine transformation. This transform rotates and scales the cubes. The rotation is limited to the z-axis to not disturb the layerwise concept of the process and the data. As the process works on a per-layer basis and the data is gathered layerwise, a rotation out of this plane should not introduce meaningful new data to the training. Hence, it is not included in the data augmentation. The spatial scaling of the data might allow the network to adapt to differently sized anomalies as it produces a more diverse training dataset with respect to the size of the anomalies. The following augmentation steps of randomly shifting the intensity and adjusting the contrast also aim at increasing the training data diversity. Together with the introduction of random Gaussian noise, this should improve the model's performance and robustness. In the last step, the ground truth label map is dilated by one voxel in each direction, similar to Section 4.5.2.

As pointed out above, only a single channel is used as input for the network. This results in six training runs, with TED^{max}, TED^{min}, TEP^{max}_{high}, TEP^{min}_{high}, TEP^{max}_{low}, and TEP^{min}_{low} as individual training inputs. Additionally, one network is trained with all six channels as input as a reference.

Validation Data The training is validated on the samples A6, A12, A13 and A30. The cropping of the 96x96x96 voxel cubes is done analogously to the training data. Additionally, the same dilation of the ground truth labels is performed. Validation is run per epoch on all validation cubes by calculating the Dice loss and coefficient. Pseudo-testing is conducted on the qualified specimens A11, B17 and B23. For this, the entire sample is used for inference with a sliding window approach. The metrics introduced in Section 4.4 are computed per epoch. The results are presented in Section 5.2.3.

Training Configuration The training parameters are kept constant for all six training runs. The models are trained using the Adam optimizer with a learning rate of 10^{-4}. The Dice loss with a softmax activation is chosen. The training batch contains four cubes randomly chosen from the training data. All six models

Table 4.5 Training data and pre-processing steps for the 3. OM CNN approach.

Training Data	
Channels	only one channel per training
Samples	A: 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29
Number of Anomalies	1154
Data Pre-Processing/ Augmentation	Intensity Scaling (0–1)
	120x120x120 voxel centered cubes
	Sampling of anomalies larger than 0.5 mm
	96x96x96 voxel random cubes
	Random Affine (rotate: 0, 0, -180–180; scale: -0.5–0.1; probability: 1)
	Random Intensity Shift (offset: -0.1–0.1; probability: 0.5)
	Random Contrast Adjustment (gamma: 0.9–1.1; probability: 0.5)
	Random Gaussian Noise (mean: 0; σ: 0.05; probability: 0.1)
	Dilate Anomalies (size: 1 voxel; connectivity: 1)

are trained for 100 epochs. The trainings are run independently from each other using monai (Version 0.9.1) and Pytorch Lightning. The results are presented in Section 5.2.3.

4.6.4 4. OM U-Net: Defect Modes

To analyze the network's generalization to different defect creation modes, the 4. U-Net is trained solely on specimens from Buildjob A. Applying the model to specimens produced with different parameter sets should allow insights into the transferability from one defect mechanism to another. In particular, the transfer from Buildjob A (Lack of Fusion) to the two parameter sets used for Buildjob B (Keyhole) is investigated. As described in Section 4.1, the parameter sets are chosen to provoke lack of fusion by increasing the scan speed (Buildjob A) or to provoke keyhole by decreasing the focus area or reducing the skywriting delay (Buildjob B). It must be noted that while the parameter sets are chosen to provoke the specific defect type, it cannot be proven that all created defects follow the desired creation mechanism.

There are at least two possible ways for the network to detect a defect. First, the network might learn to detect the creation process of the defect itself. In this case, the model would be sensitive to radiation fluctuations induced by changes in the meltpool dynamics. These dynamics are closely linked to the stability of the process and, hence, to the formation of defects. Specific defect formation mechanisms might lead to specific signatures in the meltpool data. For example, the formation of a keyhole might lead to a different meltpool signature than the formation of a lack of fusion. If this is the case, training a neural network on a large variety of defect-creation mechanisms might become necessary. This would significantly increase the effort to create the training and test dataset. Furthermore, this would limit the application of the model to known cases of defect formation mechanisms and, hence, drastically complicate the industrial implementation and qualification.

The second possible way for the network to detect a defect is based on the defect's influence on the process signatures in later stages. For example, the defect might influence the heat flux of the subsequent layers as it acts as an insulator and hinders the energy drainage by the material. The increased energy accumulation alters the heat radiation of the material in the layers above the pore, pronouncing the defect.

In this case, the neural network detects the heat signature in layers above the defect. The heat signature above a pore (either keyhole or lack of fusion) is assumed independent from its original formation mechanism as either type of pore acts as an insulator in the material. Therefore, the network should be able to detect defects created by other formation mechanisms even if only trained on one defect creation parameter.

The two hypotheses are tested by training the network on one defect creation parameter set and testing it on all available parameter sets. Additionally, an explainable AI approach is investigated in the scope of this project by Milcke [2022]. It aims to identify relevant features and channels by analyzing synthetic input patterns and the internal behavior of the CNN. The results are summarized in Section 5.2.4 and Section 6.2.

Architecture The architecture of the U-Net is kept constant to Section 5.2.2. This should allow for a well-tuned training process and a better comparison of the models.

Training Data As described above, the network is solely trained on specimens from Buildjob A. The individual specimens are listed in Table 4.6. The pre-processing and augmentation steps are adapted from Section 4.6.2 and listed in Table 4.6. These result in 2028 individual training cubes. The extensive data augmentation in Section 4.6.2 showed promising results and should facilitate the training by reducing the risk of overfitting. By reducing the risk of overfitting, the model's generalization should be improved, hence improving its transferability to new data.

Table 4.6 Training data and pre-processing steps for the 4. OM CNN approach.

Training Data	
Channels	TED^{max}, TED^{min}, TEP^{max}_{high}, TEP^{min}_{high}, TEP^{max}_{low}, TEP^{min}_{low}
Samples	A: 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29
Number of Anomalies	2028
Data Pre-Processing/ Augmentation	Intensity Scaling (0–1)
	120x120x120 voxel centered cubes
	Sampling of anomalies larger than 0.4 mm
	96x96x96 voxel random cubes
	Random Affine (rotate: 0, 0, -180–180; scale: -0.2–0.2; probability: 1)
	Random Intensity Shift (offset: 0.2; probability: 1)
	Random Contrast Adjustment (gamma: 0.9–1.1; probability: 1)
	Random Gaussian Noise (mean: 0; σ: 0.05; probability: 0.5)
	Dilate Anomalies (size: 2 voxels; connectivity: 1)

Validation Data The validation data consists of samples from Buildjob A and B. The validation pre-processing steps are analogous to Section 4.6.2 and result in 2259 validation cubes. The integration of specimens from Buildjob B allows for a first interpretation of the model performance on different defect creation parameters during training. This interpretation should be considered with care as it is based on the Validation Dice, but it can indicate overfitting already during training. The relevant performance metrics are the BUD, NUD_{400} and FP_{400} and are presented in Section 5.2.4.

Training Configuration The training is run for 100 epochs with the training setup identified as well performing in Section 5.2.2. The Epoch 53 was not saved due to a server error. The Adam optimizer with a learning rate of 10^{-4} is used for minimizing the training loss (Dice loss). Batches of four training samples are selected randomly from the training dataset. The training, validation, and testing results are shown in Section 5.2.4.

4.6.5 5. OM U-Net: Defect Size

The fifth approach investigates the generalization of the model to different defect sizes. The transferability with respect to different pore sizes is crucial for the understanding of the network capabilities. If, for example, the network does not generalize well across a range of defect sizes, it cannot be ensured that the model can detect larger defects than those included in the training data set. This would limit the use of the CNN to specific defect sizes and, at the same time, hinder the use of the POD as a relevant performance metric, as the POD requires a correlation between the size of a defect and its probability of being detected. In contrast, if it can be shown that the model is able to detect defects larger than those contained in the training dataset, it can be assumed that it generalizes well beyond the training data. To estimate the influence of the defect size distribution, the training data is reduced to an artificially small size range. The model is trained solely on defects in the range of $300\,\mu m$ to $500\,\mu m$ while being evaluated on defects up to around $1\,mm$.

Architecture The standard U-Net architecture, as described in Section 4.6.1, is used to better generalize the results.

Training Data The specimens from Buildjob A and B are used for training. The detailed list can be seen in Table 4.7. As for the previous trainings, the volumes are intensity scaled to between 0 and 1 to facilitate training. Additionally, all defects are cropped to individual smaller cubes of size 120x120x120 voxels. Of those cubes, only volumes containing a defect in the range up to $500\,\mu m$ are selected for training. This is ensured by analyzing the individual cubes and determining the largest defect contained within. The subsequent data augmentation steps are analogous to Section 4.6.2 with slight adaptions in the degree of augmentation. The detailed augmentation steps are described in Table 4.7. This results in a total of 8917 training samples.

Validation Data The validation data is produced analogously to the training data by intensity scaling and cropping (120x120x120 cubes) to smaller volumes. All cubes containing defects larger than $300\,\mu m$ are considered for validation. This results in 7875 validation samples from which random cubes of size 96x96x96 are extracted. In the last step, the defect label map is dilated by one voxel as described in Section 4.6.2.

Training Configuration The training configuration is kept constant from Section 4.6.1 and Section 4.6.2 to allow for better comparability. The model is trained for 50 Epochs. The results are presented in Section 5.2.5.

Table 4.7 Training data and pre-processing steps for the 5. OM CNN approach.

Training Data	
Channels	TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, TEP_{low}^{min}
Samples	A: 3, 4, 5, 7, 8, 9, 10, 14, 15, 16, 18, 19, 21, 22, 25, 26, 27 B: 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29
Number of Anomalies	8917
Data Pre-Processing/ Augmentation	Intensity Scaling (0–1)
	120x120x120 voxel centered cubes
	Sampling of anomalies up to $500\,\mu m$
	96x96x96 voxel random cubes
	Random Affine (rotate: 0, 0, -180–180; scale: -0.5–0.1; probability: 1)
	Random Intensity Shift (offset: 0.1; probability: 1)
	Random Contrast Adjustment (gamma: 0.9–1.1; probability: 0.5)
	Random Gaussian Noise (mean: 0; σ: 0.05; probability: 1)
	Dilate Anomalies (size: 1 voxel; connectivity: 1)

5 Results

5.1 Computed Tomography

In the following, the results of the CT CNN trainings described in Section 4.5 are presented. As pointed out previously, on the one hand, the results are used as a tool for the creation of the GT label maps for the OM CNNs. On the other hand, they represent their own field of research and contribute to the state of the art of automatic CT image analysis. The label maps and experiments are created iteratively to improve the CNN performance and label map quality at the same time. The main objective of the three approaches outlined in the following is generating a high-quality label map with reasonable effort. Therefore, the results focus on the performance evaluation with respect to POD, BUD, NUD_{400} and FP_{400}.

5.1.1 1. CT U-Net: Baseline

The results of the first training iteration for the segmentation of the CT volumes are shown in the following. The training setup and datasets are described in detail in Section 4.5.1. Firstly, the training progress and validation results are introduced. Secondly, the test results on the qualified GT are presented. Finally, the results are reviewed in short, and the next steps for improvement are discussed.

Training Results As pointed out in Section 2.3, the objective of the training process is the minimization of the training loss function. In the presented case, the loss function is defined as the Dice loss of the binary segmentation task (pore vs. background). As shown in Figure 5.1, the training loss decreases significantly in the first epochs before leveling out and fluctuating at around 0.25. The same holds true for the validation loss qualitatively but at a lower loss between 0.1 and 0.05 (Figure 5.2). In combination, both values suggest a well-performing training process in which the U-Net can quickly adapt to the given data. Additionally, the validation loss indicates a good generalization of the network beyond the training data. The validation Dice in Figure 5.3 supports this assumption as it initially increases and then fluctuates around 0.63. This can be regarded as a relatively high Dice score for a complex segmentation task. Furthermore, it indicates that the network has not yet overfitted to the training data as the validation scores do not deteriorate over the training process.

In addition to the loss and Dice score, the performance of the U-Net can also be evaluated visually. It is important to note that the label maps used for this validation cannot be regarded as qualified GT as they are created by non-qualified inspectors, partially with the help of the semi-automated workflow as described in Section 4.3.3 and Section 4.5.1. Therefore, the validation scores and images can only be used for qualitative analysis and do not allow for a comparison with other trainings. Nevertheless, they can deliver valuable information about the progress and generalization of the training.

Figure 5.4 shows two exemplary anomalies chosen randomly from the validation dataset. The CT intensity image is shown on the left, with the pore anomalies visible as darker areas in the image. The binary image in the middle shows the GT label map with the black pixels representing regions in which an anomaly is to be found. On the right, the prediction of the U-Net is depicted as a binary image. Like the GT label map, black pixels describe the area where the model detected a pore. In general, this visualization indicates that the CNN is able to adapt and generalize the underlying concept of a pore, at least in the case of this random validation sample, even if the prediction and GT label map show slight differences in the detailed structure of the pore. For further analysis, the network is applied to the qualified GT data in the following for a quantitative and comparable evaluation of its performance.

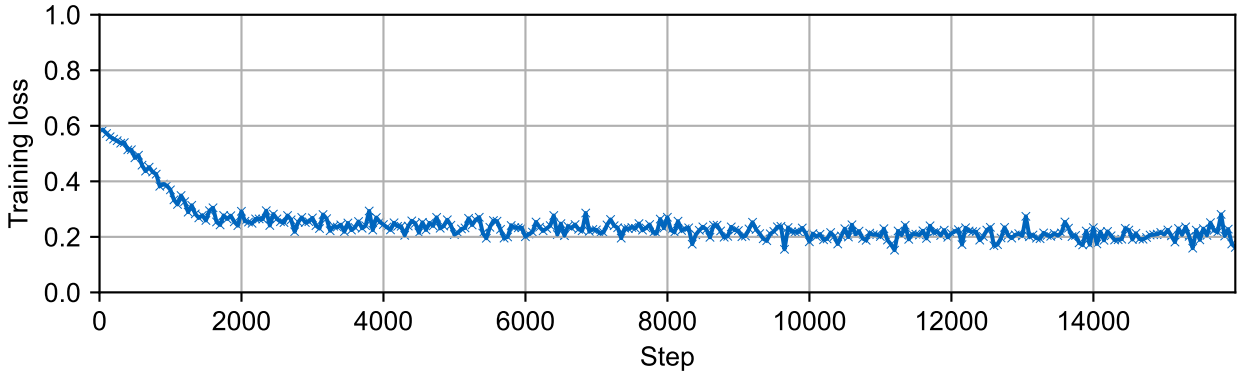


Figure 5.1 Training loss (Dice) for the 1. CT CNN trained on pores.

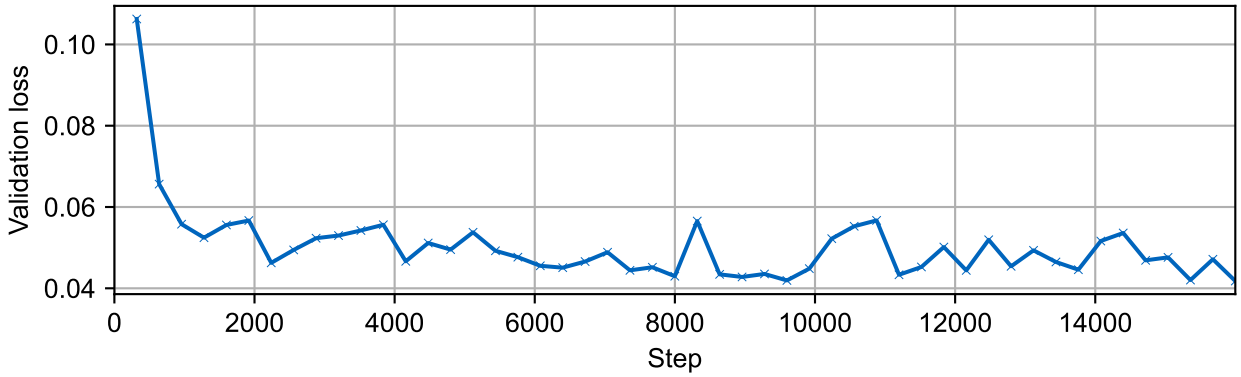


Figure 5.2 Validation loss (Dice) for the 1. CT CNN trained on pores.

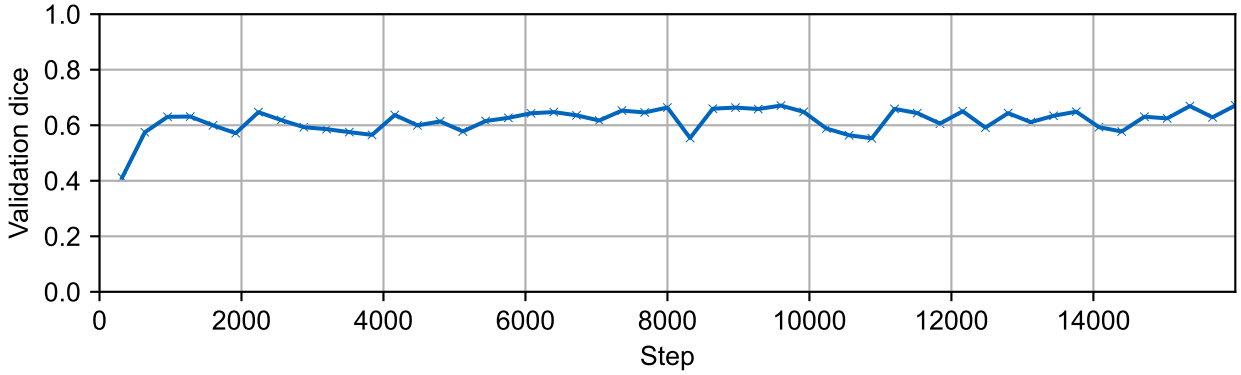
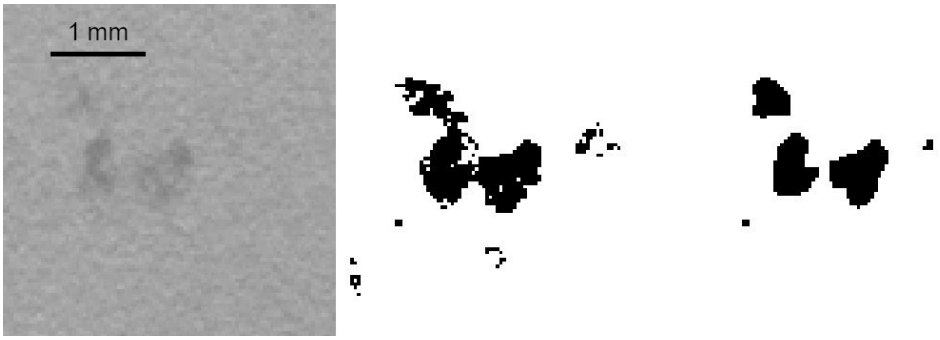


Figure 5.3 Validation Dice for the 1. CT CNN trained on pores.

(a) Prediction of the CNN at Epoch 24 on the right.



(b) Prediction of the CNN at Epoch 49 on the right.

Figure 5.4 Two exemplary anomalies in the validation dataset for two epochs (a) Epoch 24, (b) Epoch 49. With the CT intensity image shown on the left, the GT label map in the middle, and the prediction of the CNN on the right.

Pseudo-Test results For the quantitative assessment of the network performance, the qualified GT label map described in Section 4.3.4 is used. It represents the GT for the CT as it is labeled and checked by qualified CT inspectors. Therefore, it does not only allow for a quantification of individual CNN models but also the comparison of different models with each other. Here, the metrics described in Section 4.4.3 are applied. Additionally, the visual inspection of selected predictions allows for identifying possible systematic false predictions and provides additional insights into the model performance.

The quantification script, described in Section 4.4, determines whether an anomaly is found. The same script analyzes whether a prediction is a false positive or a true positive. For the quantification, only anomalies (predicted as well as GT) larger than $200\,\mu m$ are taken into consideration. Smaller anomalies are disregarded for quantification but not removed from the labelmap. The quantification focuses on sample A11, as the other qualified specimens are used as training data for later CT CNNs. A11 contains a statistically sufficient amount of anomalies and, therefore, still allows for a well-founded evaluation while allowing for a comparison with subsequent CNNs.

Epoch 49 is chosen for analysis as it shows a high Dice score. It predicts 1995 anomalies in total for A11. Of those predicted anomalies, only 220 represent actual anomalies according to the qualified label map. Hence, 1775 detected anomalies have to be regarded as false positive detections. On the other hand, the CNN detects 220 of the 228 actual anomalies. In combination with a BUD of around $465\,\mu m$, the detection performance for anomalies can be regarded as relatively high. Nevertheless, the number of false positives has to be reduced for the network to be applicable as a labeling mechanism for further studies. This is particularly true concerning the size of the largest false positive, around $896\,\mu m$. Figure 5.5 shows the POD curve for A11. The blue points show anomalies with their size marked on the x-axis and the binary class of found (100%) or not-found (0%) on the y-axis. The blue curve shows the POD curve as determined by an Matlab internal tool based on the method by Berens (see Section 4.4.2). The orange curve shows the 95% confidence bound. The relevant POD of 90/95 can be seen in the plot at the intersection of the 95% confidence bound with 90% on the y-axis. It is calculated to $0.32\,mm$, even below the largest undetected anomaly. This can be explained by the low number of false negatives and

the high number of true positives, even for small anomalies. The low POD, in contrast to a high false positive count, highlights the importance of different metrics. While the CNN performs well in detecting all defects, it does that at the cost of predicting a significant number of false positives. Hence, selecting a well-performing network should consider multiple criteria and the given use case. The metric results are summarized in the confusion matrix in Table 5.1.

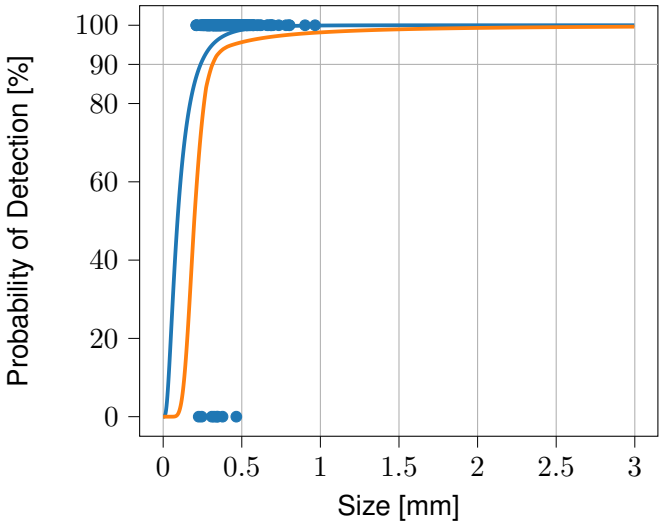


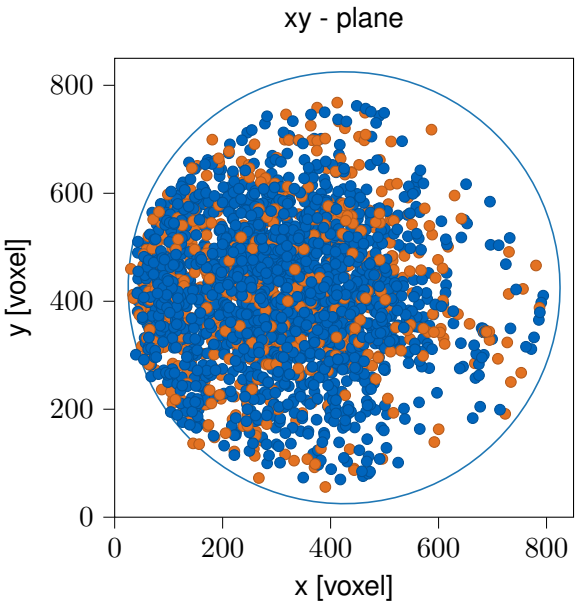
Figure 5.5 POD curve (blue) and the 95% confidence bound (orange) for the 1. CT CNN on the sample A11 resulting in an $a_{90/95}$ of 0.32 mm.

The quantitative evaluation is supported by a visual inspection of selected label maps as explained in Section 4.3.4. Figure 5.6a and Figure 5.6b visualize the appearance of false positives and true positives to manually detect possible systematics. For the inference of the 1. U-Net two minor systematics can be detected. Firstly, the falsely detected anomalies are distributed over the entire sample with a slight focus on the inner part of the specimen. This might be due to the lower contrast in the center of the sample caused partly by beam hardening artifacts in the CT scan (see Figure 5.7 inner part of specimen in the top left cross-section). Figure 5.7 also shows minor CT artifacts in regular intervals along the z-axis, which are probably caused by the specimen holder. A negative influence of these artifacts on the segmentation performance of the CT CNN cannot be observed. Secondly, an increased density of false positives within the specimen is observed around the steps encircling the specimen (Figure 5.8). Those are probably caused by CT artifacts due to the density jump at the sharp edges of the sample. Both systematics will be tackled in the subsequent experiments.

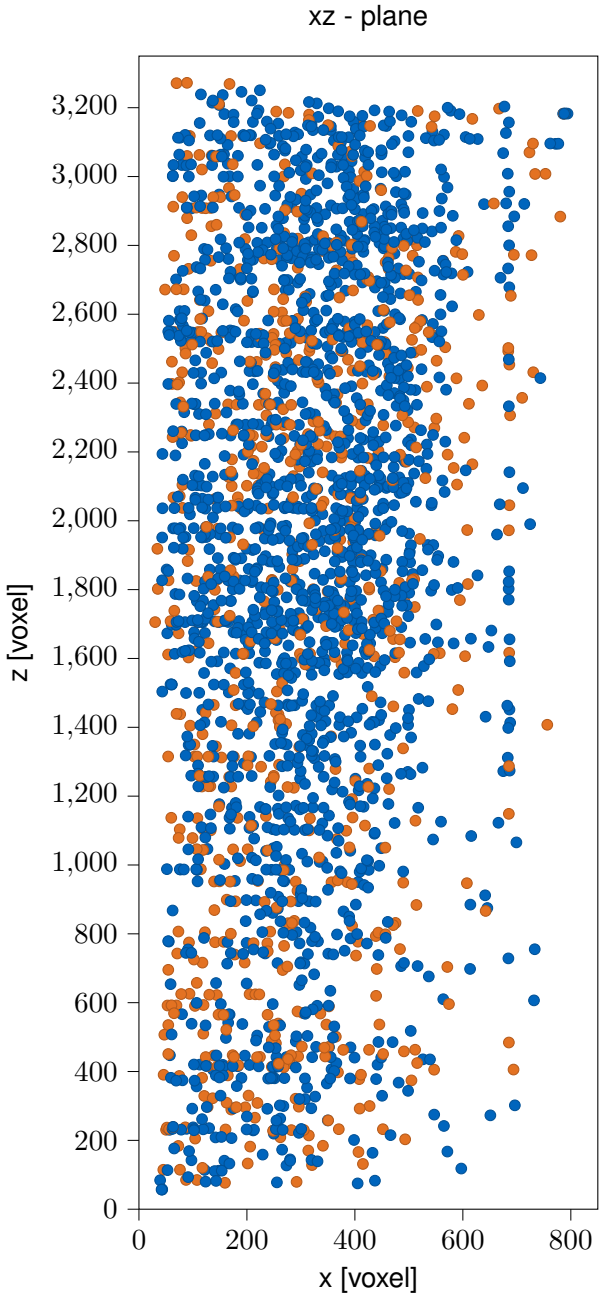
Overall, the 1. CT U-Net highlights the potential of CNNs to automatically analyze and segment CT scans. It shows a promising POD but, at the same time, a high FP count for pores. Therefore, the following experiments will focus on improving the FP count while keeping a low POD, BUD, and NUD.

Table 5.1 Confusion matrix for the 1. CT CNN.

		Ground Truth		
		Anomaly	No Anomaly	Total
Prediction	Anomaly	220	1775	1995
	No Anomaly	8	n/a	n/a
	Total	228	n/a	



(a) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 1. U-Net on the sample A11 in the xy-plane.



(b) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 1. U-Net on the sample A11 in the xz-plane.

Figure 5.6 Visualization of the predictions on A11 for the 1. OM U-Net. Due to the high number of false positives, the plot does not allow for the investigation of detailed findings. Nevertheless, it shows an increased concentration of false positives in the center of the sample.

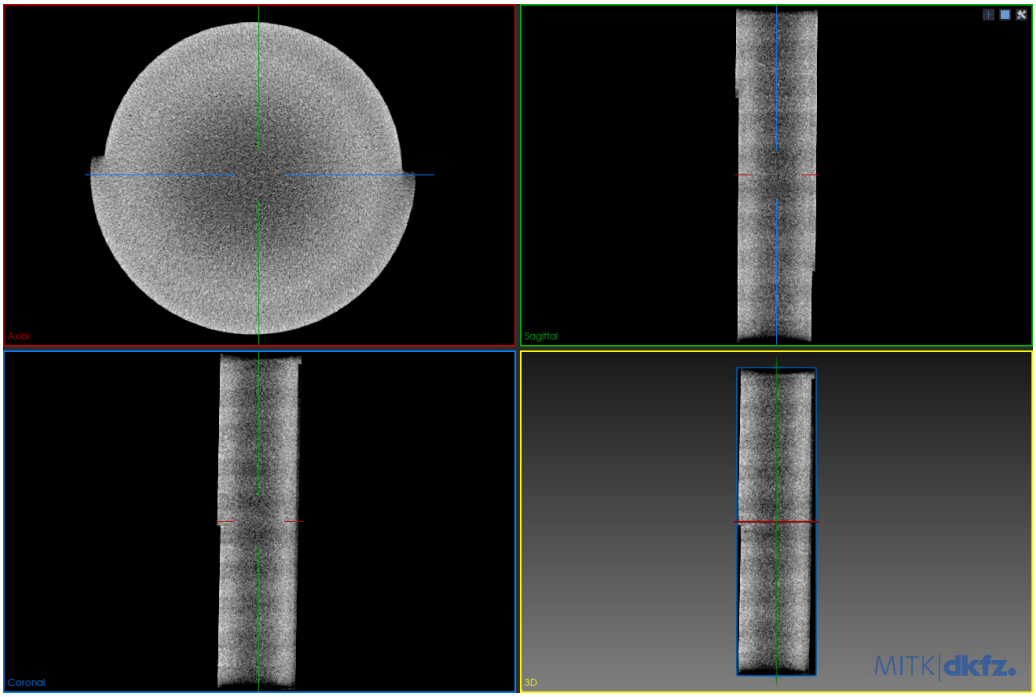


Figure 5.7 Exemplary cross-section of the CT scan of specimen A11. In the top left, the xy-plane is shown with minor beam hardening artifacts visible in the center of the sample. The top right and bottom left view show the xz- and yz-plane, respectively, with CT artifacts visible at regular intervals, which are probably caused by the specimen holder.

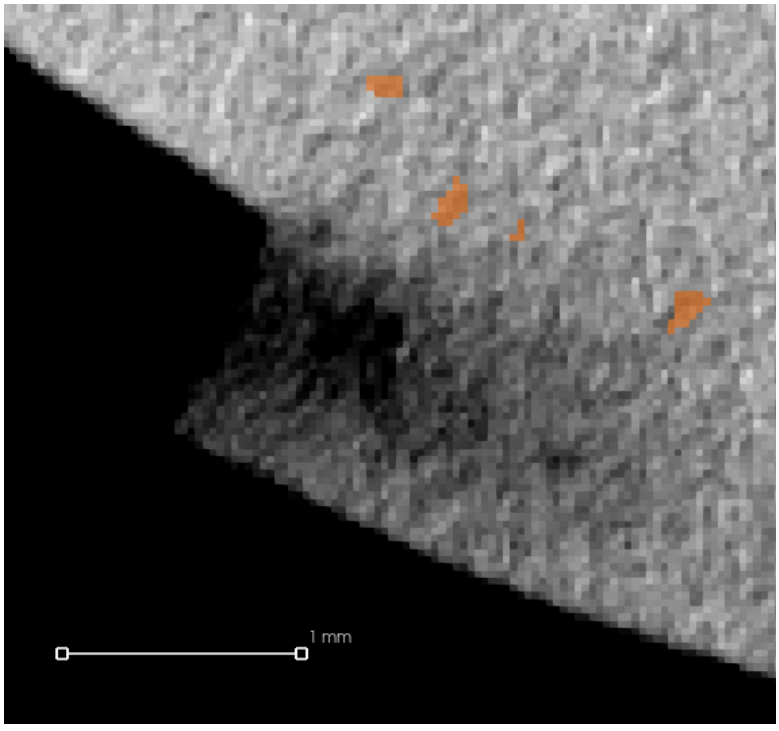


Figure 5.8 Systematic false positives (orange) around the steps encircling the specimen are probably caused by CT artifacts due to the sudden density change.

5.1.2 2. CT U-Net: Pores & Inclusions

The second CT CNN model focused on reducing false positives and incorporating inclusions into the training process. A more fine-grained presentation of the iterative steps taken to improve the model performance, including intermediate results, i.e., concerning the data pre-processing, is given in Permadi [2021]. The training configuration of the 2. U-Net is defined in Section 4.5.2.

Training results The U-Net is trained on two anomaly classes (pores and inclusions). While the training and validation loss are calculated as the mean loss of the two classes, the validation Dice score is calculated per class to allow for a better understanding of the training process. Figure 5.9 and Figure 5.10 show the training and validation loss for the 2. CT CNN. The training loss strongly decreases in the first epochs before reaching a plateau at around 0.45. At Epoch 40, a sudden drop from around 0.45 to approximately 0.2 in training loss can be observed. The validation loss decreases slowly over the entire training period with stronger fluctuations than the training loss. There is no significant drop noticeable for the validation loss. This indicates that the drop in training loss is due to the change in training data. As pointed out in Section 4.5.2, the first 40 Epochs are trained with a strong focus on pores by including approximately 32 pore patches for one inclusion patch in the training dataset. Hence, the network can adapt better to pores, which at the same time have a larger contribution to the training loss. The pores are more challenging to the network, so the training loss is higher when calculated on more pore patches. With the relative increase of inclusion patches compared to pore patches in the training data, the training loss decreases strongly, as seen in Figure 5.9. The validation dataset is not changed, and therefore, no significant drop can be seen in the combined validation loss Figure 5.10.

In contrast, the influence of the training data adaption at Epoch 40 is visible in the dedicated validation Dice score for the pores. The reduced focus on pores leads to a decline in validation Dice for this class while no clear impact for the inclusion Dice score can be detected (Figure 5.11 and Figure 5.12). This implies that the network is already able to adapt well to the inclusion data, even with a small number of samples per patch. In general, for the given CT scans, the model performs better for inclusions than pores. This can be attributed to two main factors. Firstly, the intensity difference between inclusion and specimen, as well as the difference between inclusion and surrounding, is distinct. This is due to the CT scan parameters, the specimen's material properties (Titanium), and the inclusion material (Tungsten). Secondly, the geometric variation in inclusions is much smaller, allowing the network to adapt faster. After inspecting the Dice score for both classes, Epoch 41 is chosen for further investigation as it combines the highest pore Dice score (65%) with a high inclusion Dice score (89%).

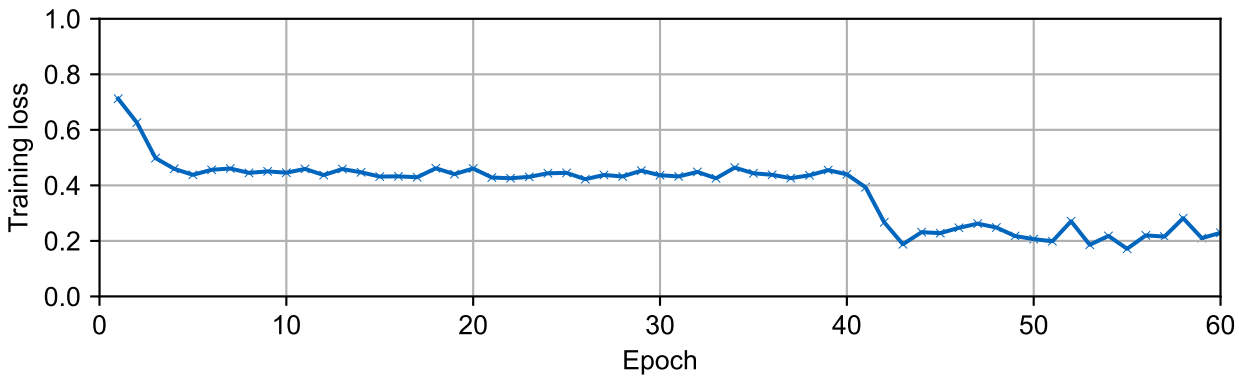


Figure 5.9 Training loss for the 2. CT CNN trained on pores and inclusions with a clear drop in loss visible around Epoch 40.

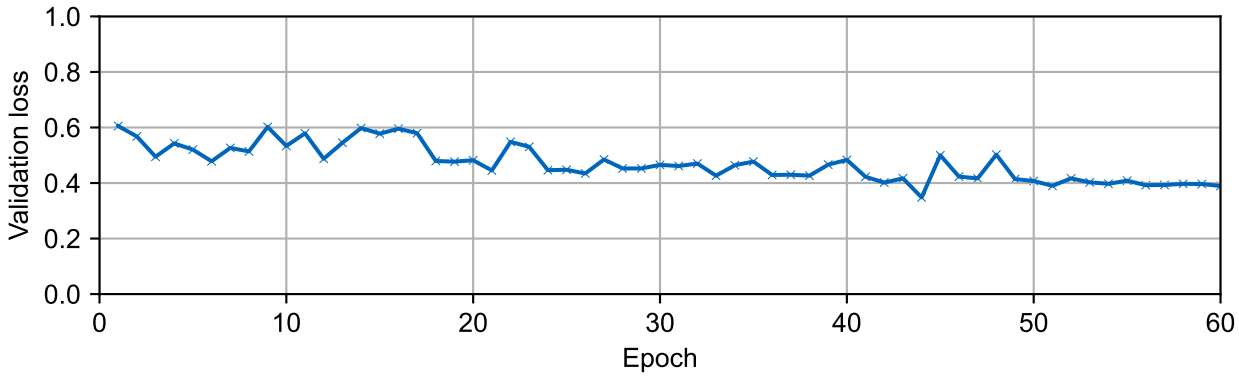


Figure 5.10 Validation loss for the 2. CT CNN trained on pores and inclusions. No clear drop is visible for the validation loss around Epoch 40.

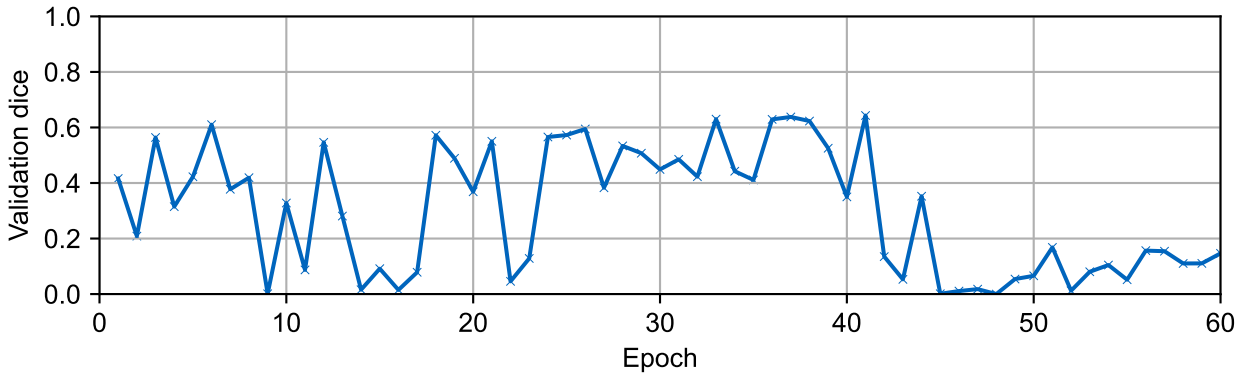


Figure 5.11 Validation Dice for pores for the 2. CT CNN trained on pores and inclusions.

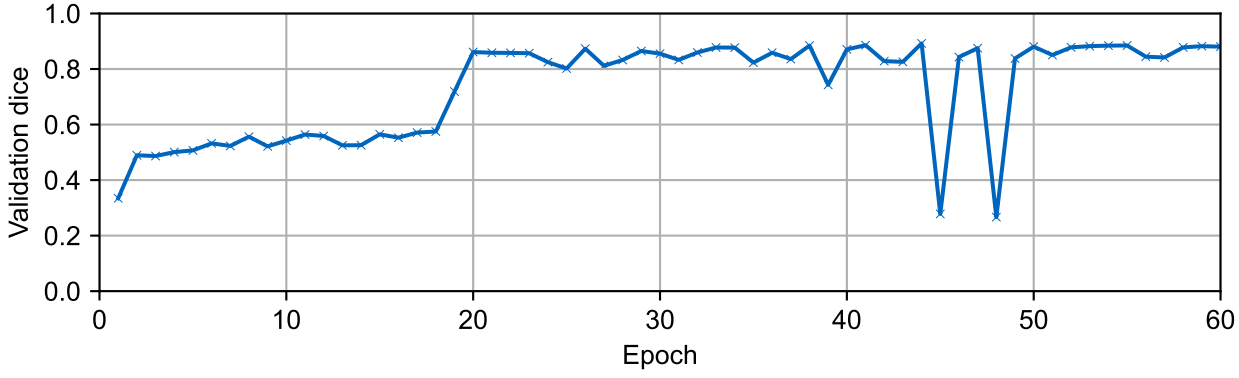


Figure 5.12 Validation Dice for inclusions for the 2. CT CNN trained on pores and inclusions.

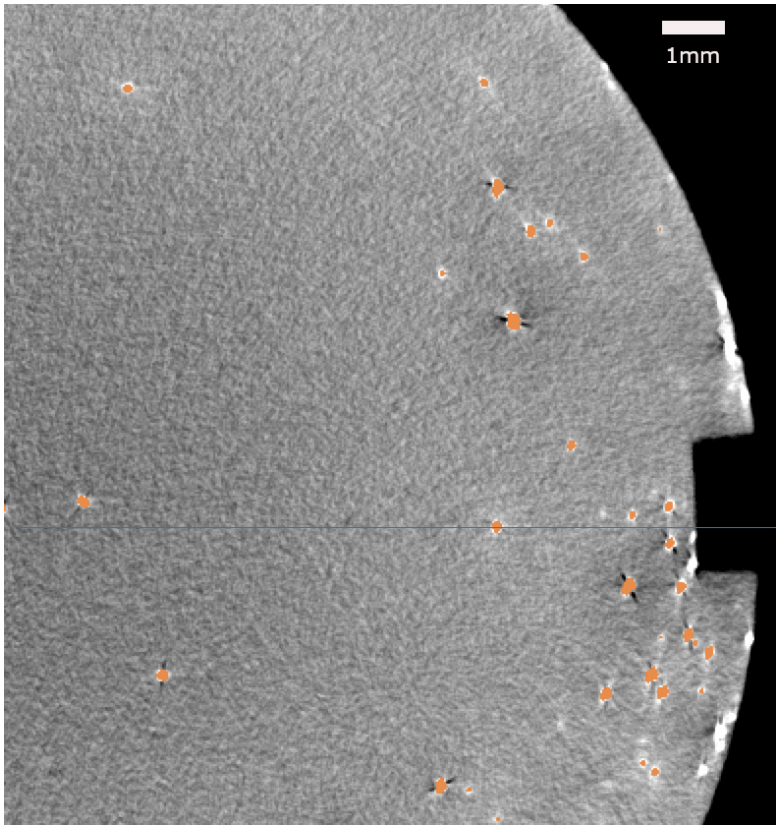


Figure 5.13 Exemplary cross-section with a high number of inclusions. The prediction label map (orange) shows a good detection performance with a slight undersegmentation per inclusion.

Pseudo-Test results The test dataset does not contain qualified CT labels for inclusions as all inclusions in the produced samples are below the threshold relevant for qualified inspection while at the same time not posing a high challenge for the CNN. Therefore, the available resources focused on labeling the more complex and challenging pores. A visual inspection of the inference label map for inclusions shows a good detection performance with a slight undersegmentation (Figure 5.13).

The performance of the 2. CT CNN on pores is evaluated analogically to the 1. CT CNN on specimen A11. A total of 955 anomalies are detected by the 2. CT CNN. Of those, 226 are true positives, and 729 are false positives. This decreases the number of false positives by around 1000 compared to the 1. CT CNN. The number of false negatives is decreased to two, with the largest not detected anomaly having a diameter of $340\,\mu m$. Hence, the network's performance could be improved regarding the number and size of false negatives. On the other hand, the amount of falsely labeled anomalies and, in particular, the size of the largest false positive of around $8.65\,mm$ needs further improvement. As Figure 5.14a shows, most false positives are located on the left side of the sample with no apparent pattern visible. Figure 5.14b does not show a clear schematic either but highlights some areas of interest, i.e., the top and bottom of the specimen as well as some specific layers.

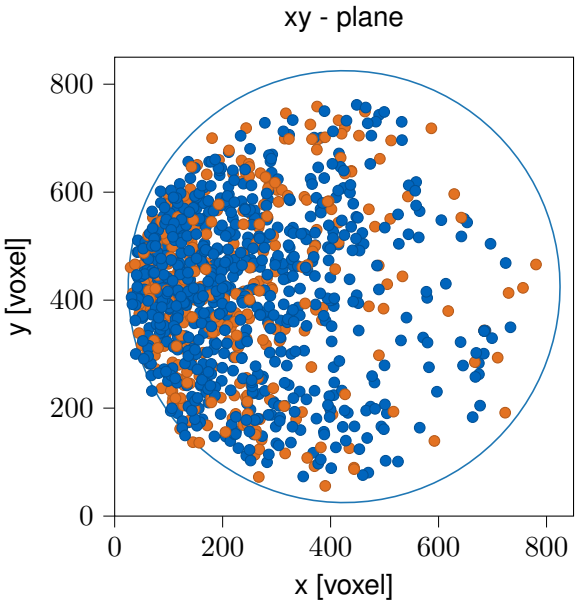
Table 5.2 Confusion matrix for the 2. CT CNN.

		Ground Truth		
		Anomaly	No Anomaly	Total
Prediction	Anomaly	226	729	995
	No Anomaly	2	n/a	n/a
	Total	228	n/a	

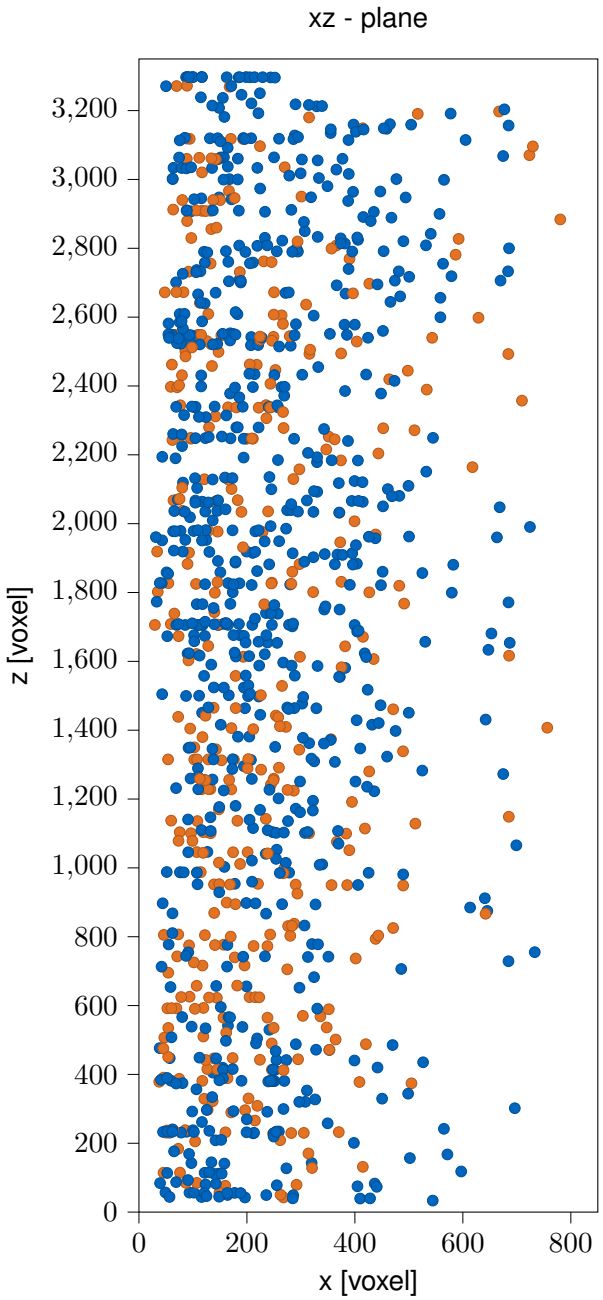
Inspecting those regions of interest shows the challenges posed by labeling CT data as described in Section 4.3.4. Figure 5.15a shows the segmentation of the 2. CT CNN in orange, and the GT label map in blue. The overlap of both label maps is visualized in purple. Figure 5.15b shows the same CT layer without overlay. The displayed slice of the CT volume contains multiple possible anomalies. Three of the possible anomalies are marked by the inspector as relevant, while the other anomalies are not relevant according to the adapted criteria. This is mostly attributed to the different sizes of the anomalies. It is important to note that the displayed cross-section only represents a fraction of the information the decision about the relevance of an anomaly is based on. In particular, a pore might increase or decrease in diameter in one of the adjacent layers. Therefore, a qualified inspector evaluates an anomaly based on the 3D data by looking through neighboring layers before deciding on the relevance of the anomaly. Additionally, it is important to note again that the qualified criteria are adapted to the task of segmentation and do not represent the actual qualified process implemented in the industry. This results in the ambiguity in the GT label maps as described in Section 4.3.4. Nevertheless, they represent the gold standard for evaluating CT scans and are therefore regarded as the reference in combination with the visual inspection.

For the 2. CT CNN, the POD could not be computed as the algorithm did not converge for the given data. This is probably due to the low number of false negatives and the high number of true positives, even for small anomalies. When inspecting the distribution of found/ not-found GT anomalies (Figure 5.16), it can be seen that the two not-found anomalies are considerably larger than the smallest found anomalies. The POD algorithm requires a connection between the size of an anomaly and the probability of it being detected. In the lack of a calculable POD, the size of the largest undetected anomaly is used as a performance indicator. With $340\,\mu m$, it is considerably smaller than for the 1. CT CNN.

Therefore, it can be assumed that the detection performance for the 2. CT CNN is qualitatively better than for the 1. CT CNN. In combination with the significantly lower number of false positives, the lower BUD also indicates a better overall performance of the 2. CT CNN, when compared to the 1. CT CNN.

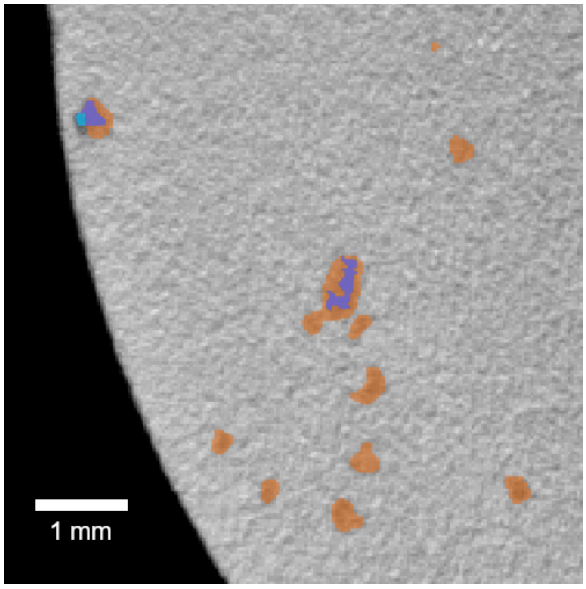


(a) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 2. OM U-Net on the sample A11 in the xy-plane.

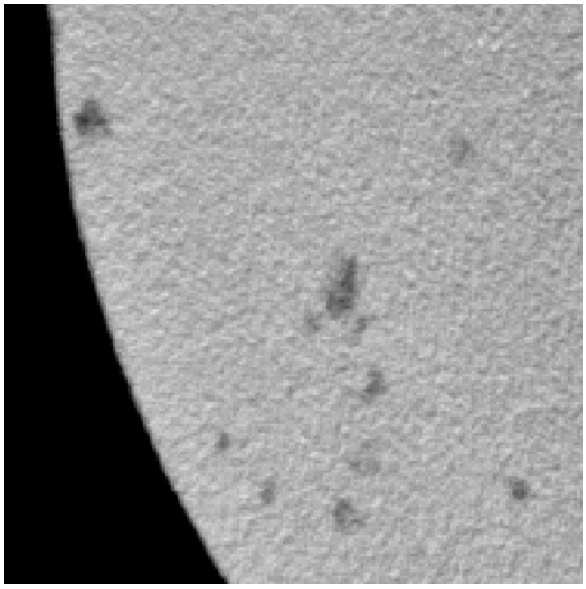


(b) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 2. UNet on the sample A11 in the xz-plane.

Figure 5.14 Visualization of the predictions an A11 for the 2. OM U-Net.

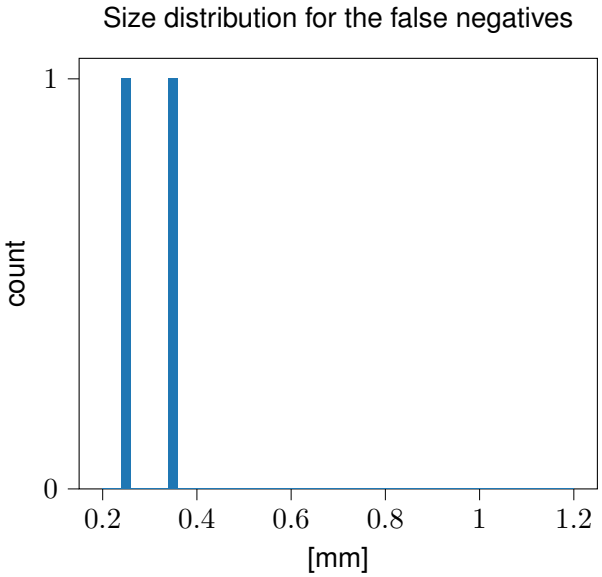


(a) CT intensity image with GT label map (blue) and prediction of the 2. CT CNN (orange). The overlap of both label maps is visualized in purple.

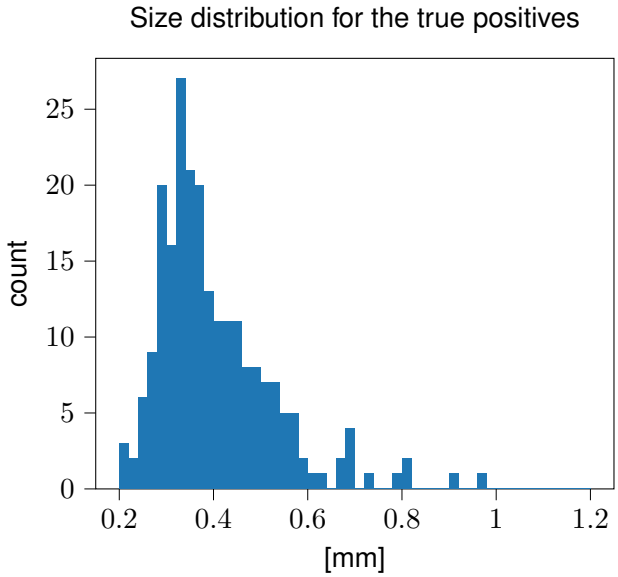


(b) CT intensity image without annotations.

Figure 5.15 Exemplary layer of the CT data of A11 in which the 2. CT CNN detected a large number of false positives. For a non-qualified inspector, the relevant anomalies are hard to distinguish from irrelevant false positives. It is important to note that the evaluation procedure is adapted to the task of voxelwise segmentation and, therefore, does not represent the procedure as qualified in the industry. This results in ambiguous labels as described in Section 4.3.4.



(a) Histogram of the size distribution of the not-found GT anomalies (false negatives).



(b) Histogram of the size distribution of the found GT anomalies (true positives).

Figure 5.16 Histograms of the false negative and true positive predictions of the 2. CT U-Net. Only two anomalies are not detected by the CNN. Both undetected anomalies are smaller than 400 µm.

5.1.3 3. CT U-Net: Performance Optimization

The 3. CT CNN incorporates qualified labels in the training, which should further improve the model performance. The training setup and datasets are described in Section 4.5.2.

Training results The training and validation loss decrease rapidly during the first training steps before leveling out at around 0.15 and 0.2, respectively (Figure 5.17 and Figure 5.18). The validation Dice for pores, on the other hand, first rises to a maximum of 0.24 before falling to close to zero. After around 2500 steps, it slowly rises back to around 0.22 over approximately 10000 steps (Figure 5.19).

One possible explanation for the sudden drop and subsequent slow recovery of the validation Dice is the adoption of the network to focus only on the newly provided qualified data. During the first training steps of the CNN, the weights are mostly influenced by the prior training in Section 4.5.2. With increasing training steps, the CNN adopts more and more to the newly provided data. The difference in the behavior between validation Dice loss and validation Dice score might be due to the different computation methods. While the Dice loss considers all classes (pores, inclusions, background), the Dice score is computed solely on the pore class. Hence, the validation Dice score is more sensitive to changes in the pore segmentation than the Dice loss. Therefore, the Dice score curve shows a stronger variation than the Dice loss curve. The qualitative behavior of both curves correlates indirectly, which is in line with the computational methods and previous results.

Pseudo-Test results Two checkpoints in the training process are used to quantify the CNN performance and interpret the results compared to the other networks. The first checkpoint is taken at Epoch 2 (Step 395), as it shows the highest validation Dice score. The second checkpoint (Epoch 95, Step 12670) is chosen to investigate the influence of the new data on the network performance and get more insight into the above-mentioned training behavior.

For **Epoch 2**, the model detects 305 anomalies in total. 217 of those are real anomalies, while 88 are regarded as false positives. This is a further improvement to the 2. CT CNN (729 down to 88) for the detection of false positives. On the other hand, the number of false negatives did increase from 2 to 11. While false negatives can be regarded as more severe to the model performance than false positives, the size of the missed anomalies must be considered. Figure 5.20 shows the size distribution of all missed (false negative) and found (true positive) anomalies. The biggest undetected anomaly is 378 µm in size, with the other undetected anomalies having a size between 200 µm and 340 µm. This results in a POD of 0.33 mm (Figure 5.22). Hence, the 3. CT CNN drastically reduced the number of false positives compared to the two previous CNNs while maintaining a low POD and false negative count. The visual inspection of the predictions finds that the schematics present in the 1. CT CNN are no longer visible in the specimen (i.e., false positives around the steps of the specimen and at the top and bottom). This is also supported by

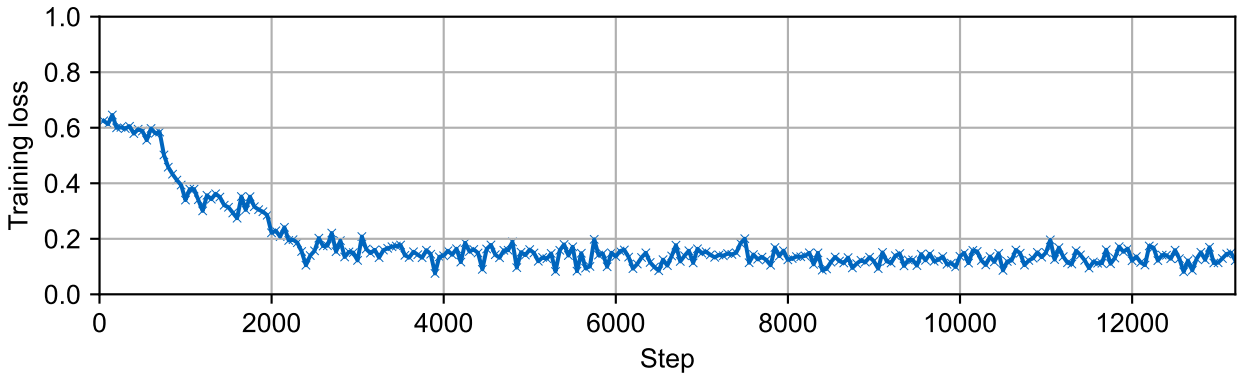


Figure 5.17 Training loss for the 3. CT CNN trained on pores from the qualified label maps.

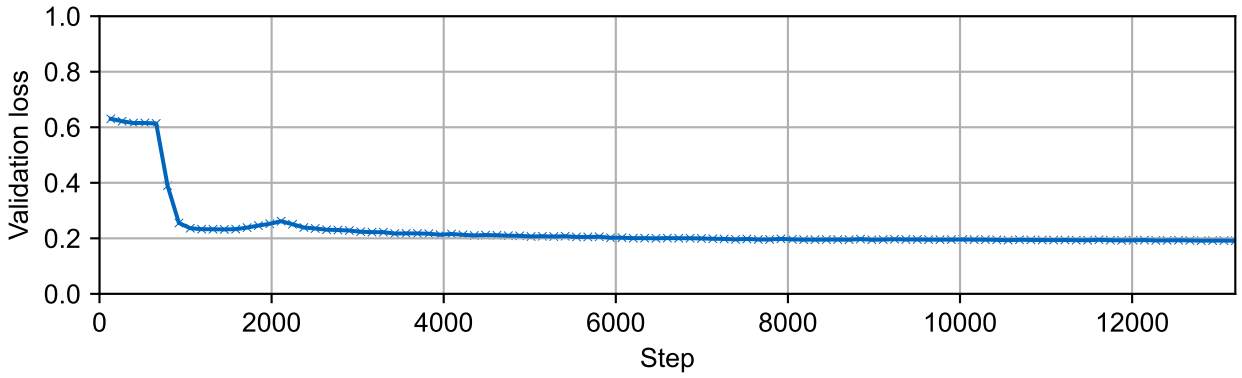


Figure 5.18 Validation loss for the 3. CT CNN trained on pores from the qualified label maps.

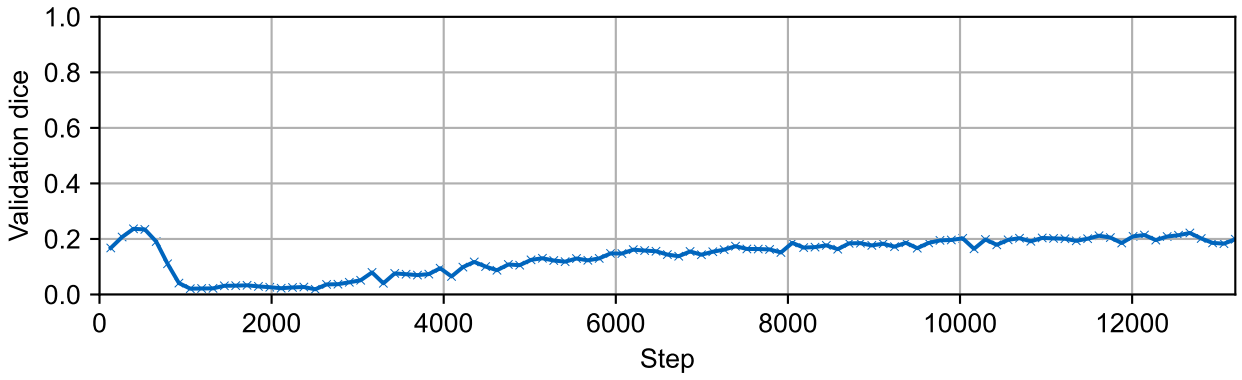
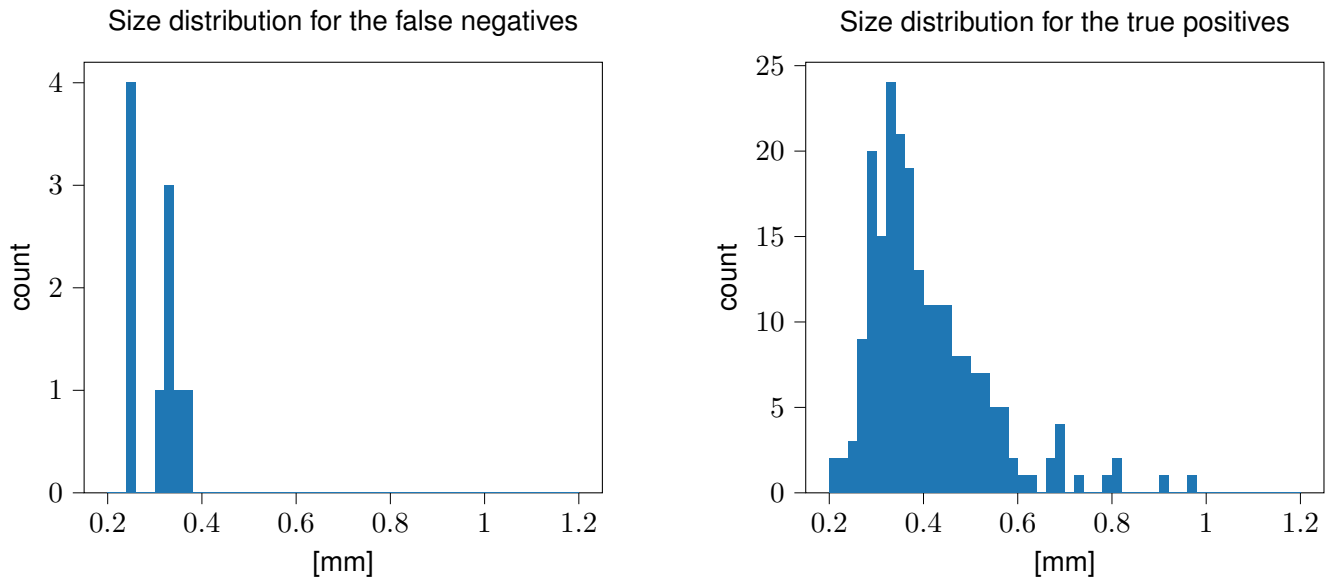


Figure 5.19 Validation Dice for the 3. CT CNN trained on pores from the qualified label maps and validated only on pores.



(a) Histogram of the size distribution of the not-found GT anomalies (false negatives).

(b) Histogram of the size distribution of the found GT anomalies (true positives).

Figure 5.20 Histograms of the false negative and true positive predictions of the 3. CT U-Net. The CNN does not detect eleven anomalies. All undetected anomalies are smaller than 400 µm.

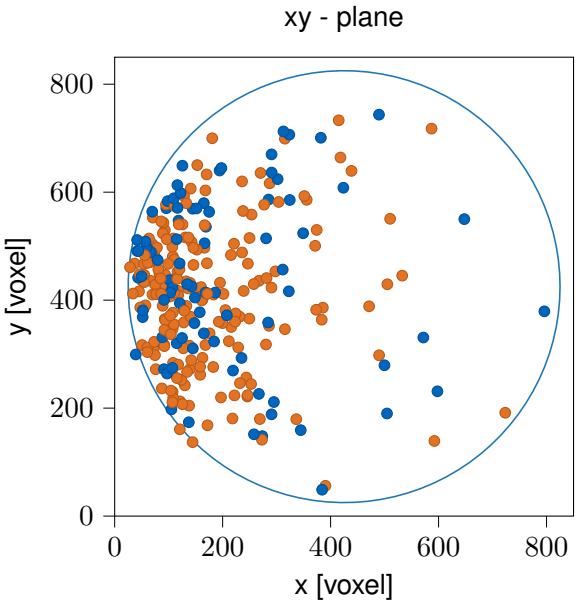
the plots in Figure 5.21a and Figure 5.21b which do not show a clear pattern for the false positives. Both findings indicate that the network is able to adapt to the qualified label maps introduced to the training dataset.

Table 5.3 Confusion matrix for the 3. CT CNN Epoch 2.

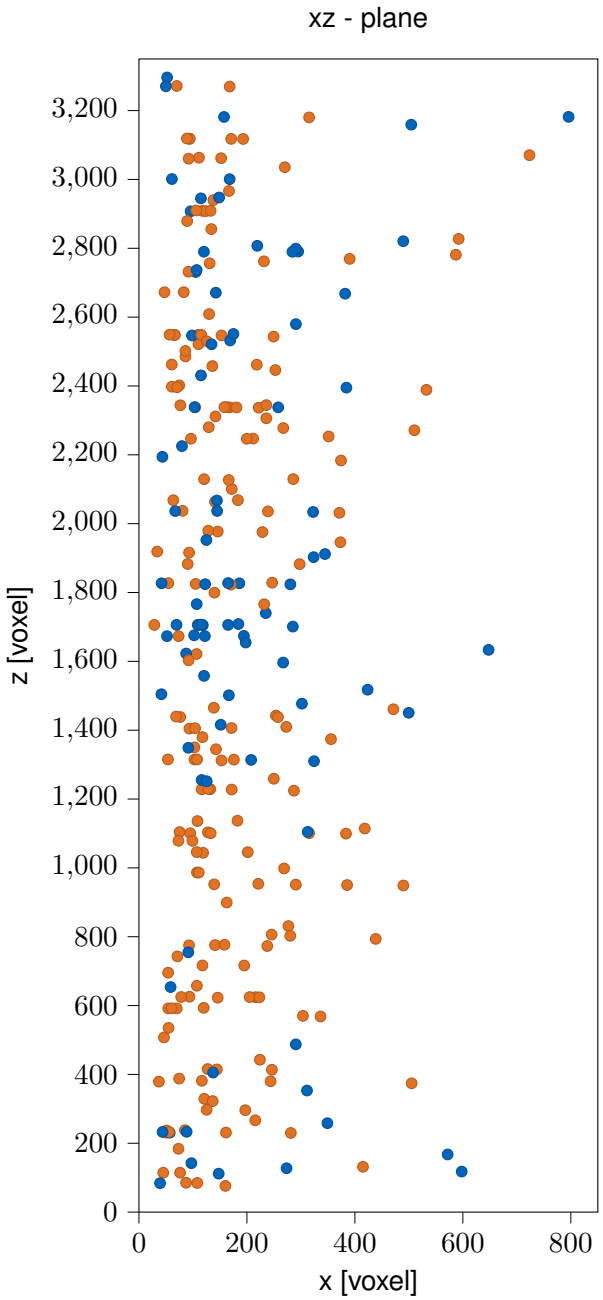
		Ground Truth		
		Anomaly	No Anomaly	Total
Prediction	Anomaly	217	88	305
	No Anomaly	11	n/a	n/a
	Total	228	n/a	

For the **Epoch 95**, the detection performance of the model degraded drastically. There are 85 false negatives, with the largest undetected defect being around 555 µm. The POD is calculated to 0.57 mm, significantly higher than the previous CT CNNs. Therefore, while the number of false positives drops to only three, the model performance is considered inferior to Epoch 2 due to the significant decrease in detection performance.

Overall, the finetuning of the 2. CT CNN on parts of the qualified label maps further improves the CT CNN performance. The combination of the high detection performance of the 2. CT CNN, with the incorporation of new qualified labels, leads to the desired reduction of false positives. In particular, this behavior is visible in the early epochs of the training. The good performance for Epoch 2 and the observed performance degradation for Epoch 95 suggest that a gradual finetuning within a few epochs is beneficial and sufficient for performance improvement.



(a) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 3. CT U-Net on the sample A11 in the xy-plane.



(b) Visualization of all true positives (orange) and all false positives (blue) for the prediction of the 3. CT U-Net on the sample A11 in the xz-plane.

Figure 5.21 Visualization of the predictions an A11 for the 3. CT U-Net.

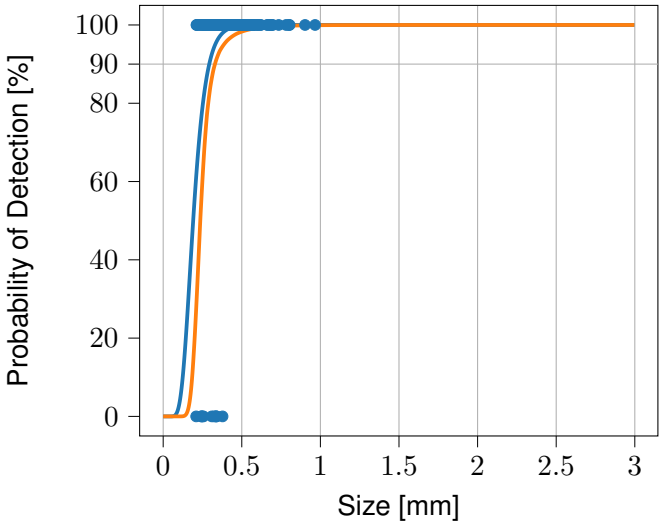


Figure 5.22 POD curve (blue) and the 95% confidence bound (orange) for the 3. CT CNN on the sample A11 with an $a_{90/95}$ value of 0.33 mm.

5.1.4 CT CNN Summary

The three presented CT CNNs are developed iteratively. In particular, the segmentation results of the previous model are used to refine the GT label maps for the next CNN training. Additionally, the network architecture and training strategies are enhanced over the iterations. Therefore, it is expected that the network performance improves over time. The results are summarized in Table 5.4. The low POD of 0.33 mm for the 3. CT CNN Epoch 2, combined with the low false positive count, is regarded as sufficient for further processing new CT specimens. The segmentation performed by the network will be used as input for the training and validation of the online monitoring models in the following. As it is not feasible to manually check all created label maps, the resulting predictions are evaluated via the qualified label maps. This is a trade-off between the affordable effort and the accuracy of the produced label maps. The influence of the CT CNN on the training of the OM CNNs will be discussed in Section 5.2.

Table 5.4 Summary of the test results for the three selected CT CNNs.

	True Positives	False Negatives	False Positives	POD [mm]	BUD [µm]
1. CT CNN	220	8	1775	0.32	465
2. CT CNN	226	2	729	n/a	340
3. CT CNN	217	11	88	0.33	378

Besides generating GT labels for the subsequent training of the OM CNNs, the presented results show the general feasibility of CNNs to automatically analyze and segment pores in CT scans. The U-Net architecture can be successfully trained to detect pores and inclusions in CT scans with CT artifacts such as beam hardening or scatter artifacts. Reducing false positive predictions in artifact-rich regions poses the main challenge, which can be mitigated efficiently by incorporating manually labeled, qualified label maps. The ambiguity of the qualified label map, described in Section 4.3.4, poses another challenge for the training and evaluation of the developed networks. This is mitigated by a visual inspection of the predicted labels. Overall, the trained networks, the iterative enhancement of the labels, and the qualified GT label maps allow for a reliable and efficient evaluation of large CT datasets. The next steps to transfer these results from research to industry are described in Chapter 6 and Chapter 7.

5.2 Online Monitoring

In the following, the results of the CNN training for the detection of anomalies based on the online monitoring data are presented. The model, the training setup, and the motivation for the respective approaches are described in Section 4.6. The general concept follows the lines of Section 5.1. Firstly, the individual model performances are presented by investigating the training progress based on the training and validation scores. Secondly, the use-case-specific test performance is evaluated based on the qualified GT data. Thirdly, specific epochs are selected for further analysis and interpretation. This allows for an in-depth understanding of the network performance and its limitations. This evaluation strategy is performed for each of the following approaches. The presented approaches represent a selection of meaningful CNNs, as listing all conducted experiments would exceed the scope of this thesis. The first two approaches show the general feasibility of the proposed method and aim to optimize the model's performance. The objective of the subsequent approaches is to better understand and evaluate the trained models and their limitations. Here, particular focus is placed on the models' transferability, robustness, and explainability. The overall results of the five approaches are summarized in Section 5.2.6. The results and their impact on future research are discussed in Chapter 6 and Chapter 7.

5.2.1 1. OM U-Net: Baseline

The first approach of the OM U-Net aims to create a baseline to show the feasibility of the presented approach to detect defects with a CNN based on online monitoring data. It uses the data of two buildjobs (A and B) as an input and is trained on all six channels (TED^{max}, TED^{min}, TEP^{max}_{high}, TEP^{min}_{high}, TEP^{max}_{low}, TEP^{min}_{low}). The exact setup and training configuration are described in Section 4.6.1.

Training results The plot of the training and validation loss (Figure 5.23 and Figure 5.24) show a quickly decreasing training loss which asymptotically approaches 0.5 before dropping to around 0.4 at step 2500. After that, the training loss oscillates around 0.4, with no significant improvements visible. The validation loss starts at a significantly lower value of approximately 0.15. It is tracked for each epoch and decreases to around 0.11 within the first epochs (first 2500 steps), similar to the training loss. A much lower oscillation is visible for the remaining training steps, with the validation loss staying around 0.11. Both losses do not improve significantly over further steps, which suggests that a local optimum is reached within the training duration of 28 epochs. The fact that the validation loss does not increase over the training implies that no substantial overfitting to the training data occurs.

The validation Dice in Figure 5.25 supports the assumption of a quickly adapting network within the first couple of epochs. It increases sharply to around 0.1 within four epochs and then levels out. At Epoch 20, a peak of around 0.16 is visible. As pointed out before, the training and validation metrics are only used to monitor the training process. The task-specific performance will be evaluated based on the POD, BUD, NUD_{400} and FP_{400} in the following.

Pseudo-Test results The network performance is evaluated using the qualified ground truth label maps introduced in Section 4.3.4. The predictions of the CNN are compared to the corresponding ground truth by the Hit-Miss analysis as described in Section 4.4.3. This is done for the samples A11, B17 and B23. Hence, all three provocation parameter sets are included in the pseudo-test set. The quantification of the CNN performance is based on the use-case-specific metrics defined in Section 4.4.

Firstly, the POD is calculated per epoch for each of the three qualified specimens (Figure 5.26). Epochs for which the POD could not be calculated are marked by NaN values. As can be seen, the POD calculation is not possible for every epoch or specimen due to the algorithm not converging for the given hit-miss results. The most common reason for the POD algorithm not converging is an insufficient data basis due to an ill-performing CNN. In such cases, the BUD and NUD_{400} offer additional insights into the model performance. The analysis of Figure 5.26 shows that no POD can be calculated for the first four epochs as the algorithm did not converge. This indicates an inferior model performance and is in line with the findings

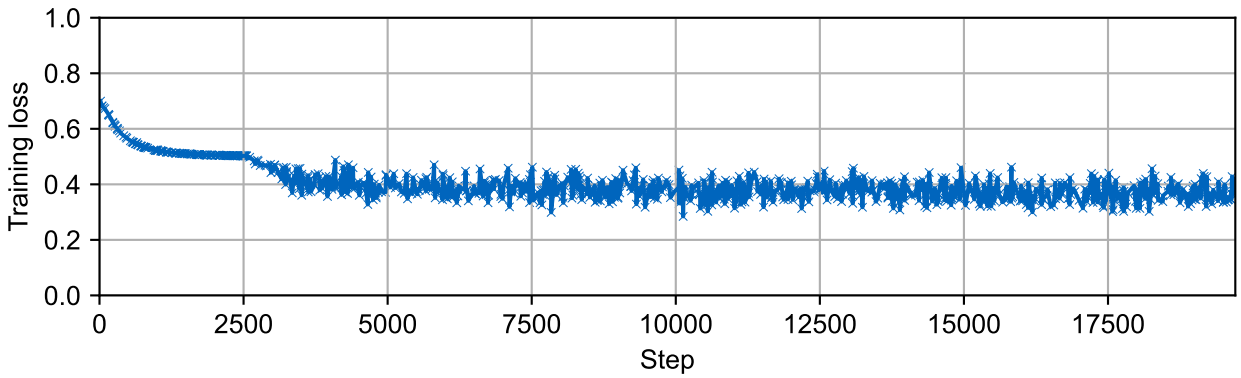


Figure 5.23 Training loss for the 1. OM CNN per training step.

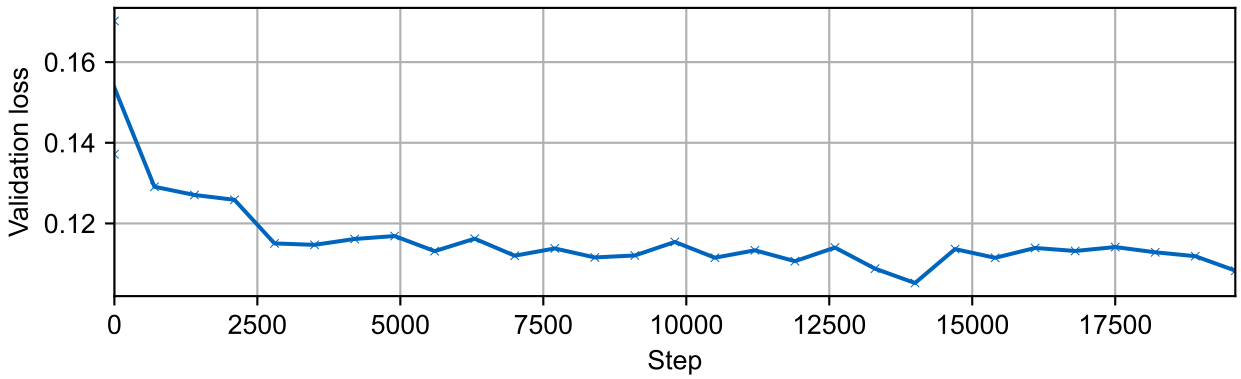


Figure 5.24 Validation loss for the 1. OM CNN per training step.

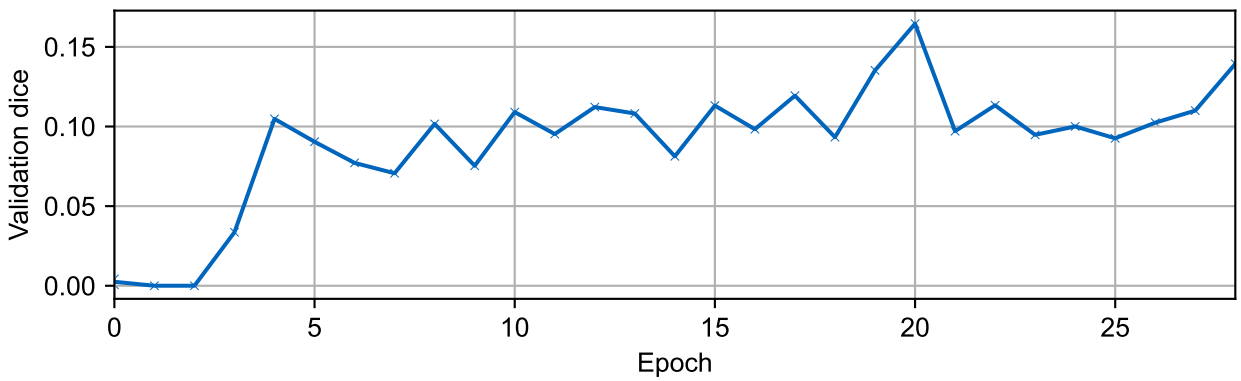


Figure 5.25 Validation Dice score for the 1. OM CNN per training epoch.

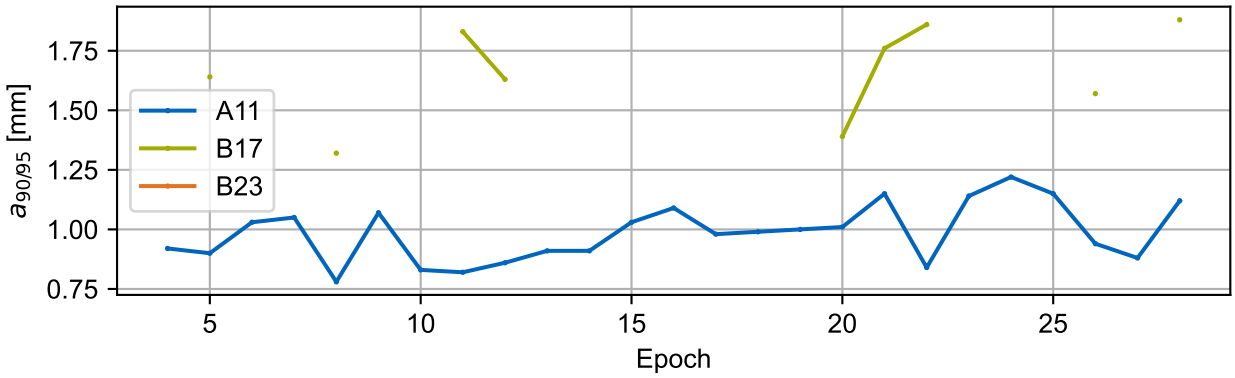


Figure 5.26 Probability of Detection for the 1. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange, not visible).

of the training results above, for which the validation Dice score only improved for the later epoch. It can be assumed that the CNN has not yet adapted sufficiently to the training data for the first epochs and is, therefore, not yet able to make meaningful predictions. This is an expected behavior for neural networks as the adaptation of weights is an iterative optimization process. For the subsequent epochs, the POD for A11 varies between 0.78 mm and 1.22 mm while the POD for B17 varies between 1.32 mm and 1.88 mm while not being calculable for some epochs at all. For specimen B23, the POD cannot be calculated for any epoch. Therefore, the additional metrics BUD, NUD_{400}, and FP_{400} will be evaluated, together with a visual inspection in the following to get a better insight into those epochs.

Figure 5.27 shows the BUD per epoch for the three qualified specimens. For the first three epochs, the BUD is equal to the largest defect within each sample. Therefore, it must be assumed that the network cannot detect relevant pores in this early stage. This confirms the assumption that the network has not adapted to the training data in the early training stage. At Epoch 3 (respectively Epoch 4 for B17), a drop in BUD is visible. For A11 and B23, the best BUD value of 782 µm and 529 µm, respectively is reached. The BUD for B17 drops to 622 µm and in a further epoch (Epoch 5) to 587 µm, which is the lowest value for B17. In the following, the BUD for A11 stays constant at 782 µm while the BUD for B17 and B23 increases to around 900 µm before jumping between approximately 600 µm and 750 µm and 908 µm respectively. The strong jumps can be explained by the design of the metric, as it only considers the maximal anomaly size. This results in a strong influence of single anomalies. For a later qualification of the system, this focus on the worst-performing prediction is benefical, as it has to be ensured that the system does not miss relevant anomalies. For the development of the CNN, additional information to evaluate the model performance is provided by the NUD_{400} and FP_{400}.

Therefore, in the third step, the number of undetected defects larger than 400 µm is taken into account. Figure 5.28 shows the NUD_{400} in percentage for each sample and epoch. For the first three epochs, the graphs show similar behavior to the BUDs in Figure 5.27. This seems reasonable, as the network cannot produce meaningful label maps due to insufficient training. From Epoch 3 (respectively Epoch 4) on the NUD_{400} drops drastically, which aligns with the results for the POD and BUD. For the later epochs, the NUD_{400} shows a more refined behavior as it focuses not solely on the largest undetected defect but the number of relevant defects. The lowest NUD_{400} for each sample is 20% for A11 at Epoch 5, 15% for B17 at Epoch 20 and 10% for B23 at Epoch 3. Inversely, this means that 80% (A11), 85% (B17) and 90% (B23) of all defects larger than 400 µm are detected.

Besides the size and amount of undetected defects, the number of false positives provides an important insight into the model performance. In Figure 5.29, the number of false positive predictions larger than 400 µm is plotted logarithmically. It shows a clear spike at Epoch 3 for all three samples, with over 1000 false positives for samples A11 and B23. In the subsequent epochs, this number decreases strongly. The lowest number of false positives for each specimen is reached at 0 FP (Epoch 13) for A11, 6 FPs (Epoch 13 and 24) for B17, and 5 FPs (Epoch 16) for B23. The lowest false positive count overall is achieved at

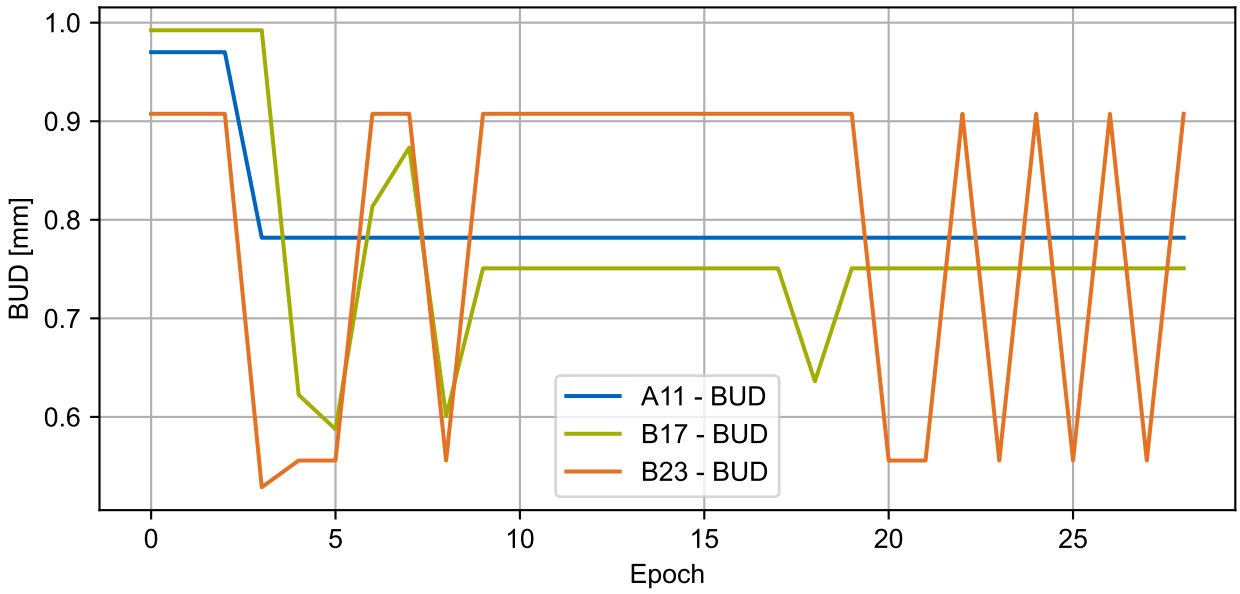


Figure 5.27 Biggest undetected defect for the 1. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange).

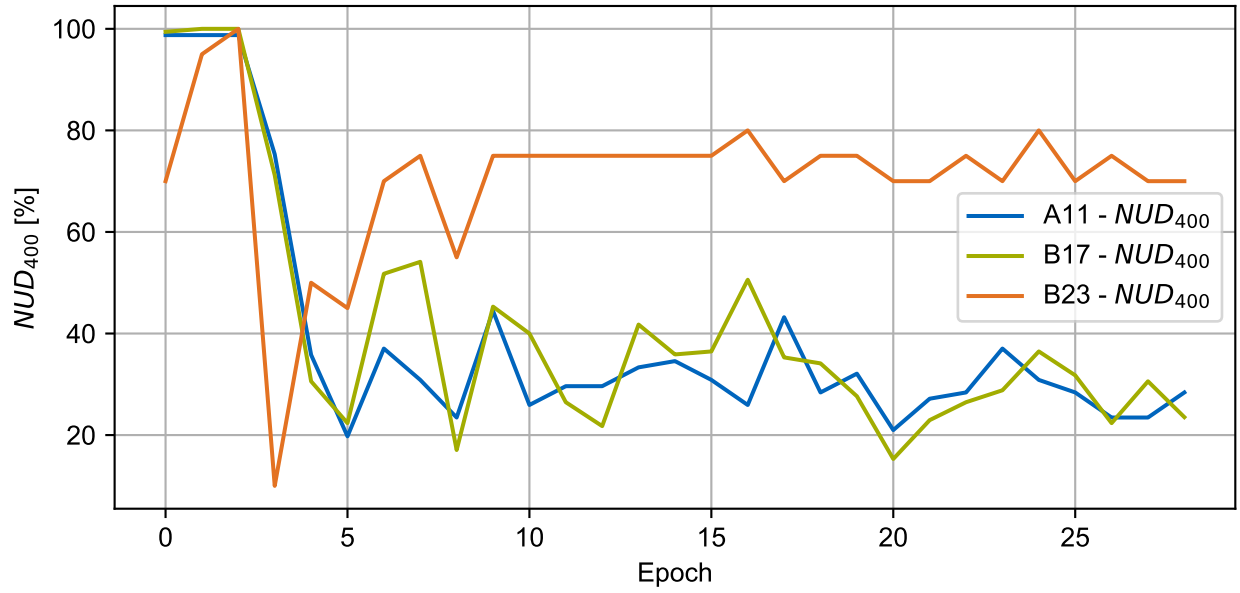


Figure 5.28 Percentage of undetected defects for the 1. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange).

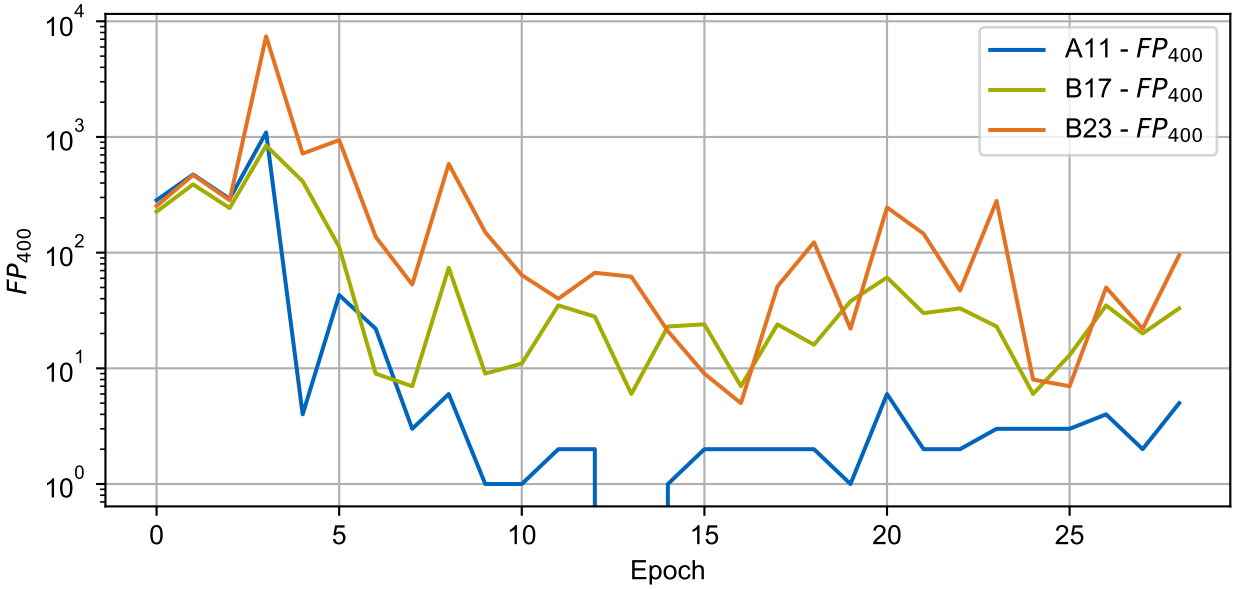


Figure 5.29 Number of False Positives larger than 400 µm plotted logarithmically for the 1. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange). For Epoch 13 there are no false positives larger than 400 µm for the sample A11.

Epoch 16, with 14 false positives larger than 400 µm. In all but two epochs (Epoch 3 and 6), the model performs worse (concerning false positive predictions) for specimens B17 and B23 in comparison to A11. In particular, B23 shows a high number of false positives, which, for most epochs, lies significantly above the number of actual relevant defects in the specimen. This behavior will be investigated in detail for the selected epochs later.

With the false positive count taken into consideration, a trade-off between false negatives (undetected defects) and false positives becomes visible. Epochs with a low BUD and NUD_{400} show increased false positives and vice versa. Figure 5.30 illustrates this behaviour exemplary for specimen A11. For Epoch 5, the NUD drops to 20 % while the false positive count rises to 45. Hence, the network is able to detect more relevant defects but, at the same time, marks more areas that are not real defects. This trade-off is common for machine learning. It depends on the later use case of the model which side should be emphasized more strongly (FN or FP). This is discussed in detail in Section 6.2.

Interpretation Combining the different metrics allows for better insights into the model performance and highlights specific epochs and behaviors for further investigation. Before focusing on specific epochs, it is important to note that the model shows a differing performance for the three specimens. While A11 consistently shows the best POD for all epochs, the POD for B17 strongly depends on the selected epoch. B23 cannot be analyzed by the POD at all. For the NUD_{400}, A11 and B17 both show the same trend over the epochs, while the model misses a significantly larger portion of defects larger 400 µm for B23. The inferior performance for B23 is also visible in the significantly higher number of false positives for B23. To further analyze this behavior and for a more in-depth analysis of the specific models, Epoch 8 and Epoch 20 are investigated in detail.

Epoch 8 is chosen for evaluation as it shows the lowest POD for A11 (0.78 mm) and B17 (1.32 mm), as well as a drop in NUD_{400} for all three specimens and a drop in BUD for B17 and B23.

There are 20 undetected defects larger than 400 µm in specimen A11 for the model at Epoch 8. The BUD (782 µm) for A11 is shown in Figure 5.31. The BUD is the same for all the subsequent epochs as well. It can, therefore, be assumed that it represents a challenging defect for the network to detect. The visual inspection shows that the label only extends over two slices in z-direction. Such labels do not

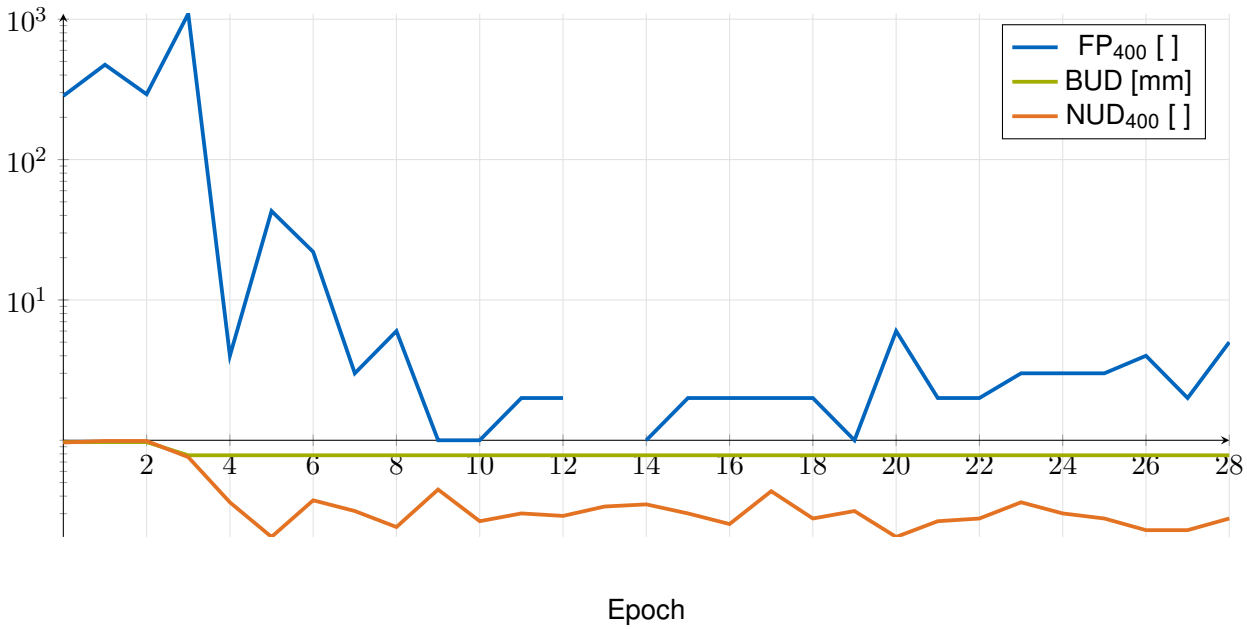


Figure 5.30 The number of false positives larger than 400 µm (blue), the biggest undetected defect (green), and the number of undetected defects larger than 400 µm (orange) for the sample A11 per epoch. For Epoch 13 there are no false positives larger than 400 µm.

have to be labeled by the inspector according to the adapted qualified procedure. Nevertheless, if the CT inspector deems a defect relevant, it can be marked even if it extends only two slices. Hence, this introduces a degree of ambiguity in the training and test data as described in Section 4.3.4. This ambiguity might impair the detection performance of the OM CNN. The same holds true for the detection of false positives. The ambiguous labeling in the ground truth leads to an increased number of false positives as the model detects anomalies that might represent pores but are not labeled as defects in the ground truth because they only extend over two layers. As described in Section 4.3.4, this ambiguity is inherent to the defined procedure. It should be considered when evaluating and comparing the CNN performance by the introduced visual inspection.

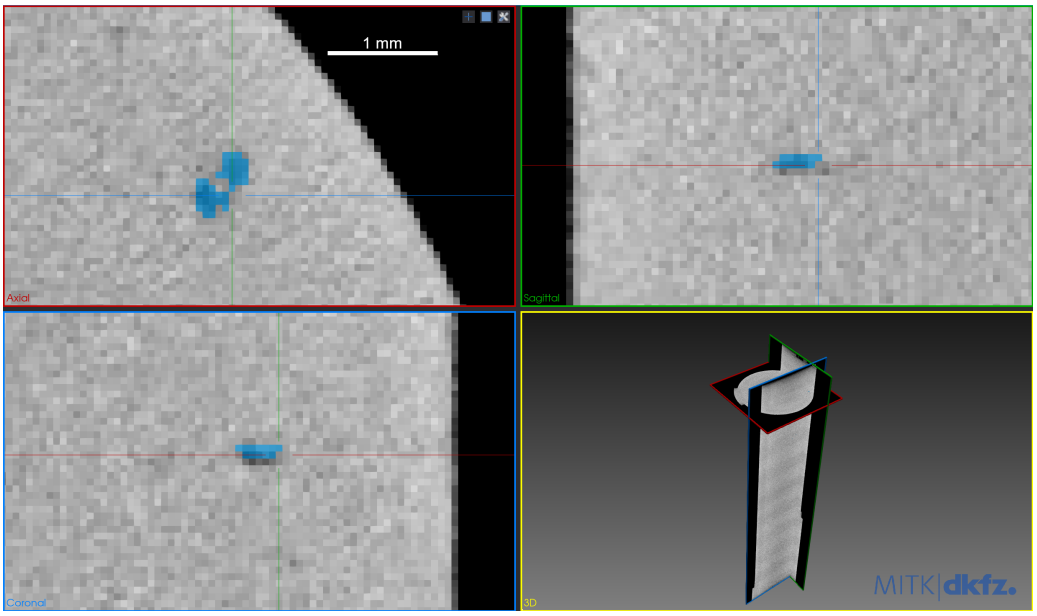


Figure 5.31 CT image of the BUD of the sample A11 with the ground truth labels marked in blue.

For B17, there are 29 undetected defects larger than 400 µm, of which the largest is 601 µm. This means that of 172 defects, 143 are found, with the largest 42 defects being detected. When manually inspecting the inference result, a good spatial correlation between the prediction and the actual defect is visible (Figure 5.32). The BUD can be seen in Figure 5.33 and shows a similar characteristic to the BUD in A11. It extends over only two slices in the z-direction and can, therefore, be regarded as a difficult, ambiguous case. The FP_{400} for B17 is 74. A visual inspection shows a mix of definitive false positives and possible ambiguous labels. No clear pattern for the FPs is observable.

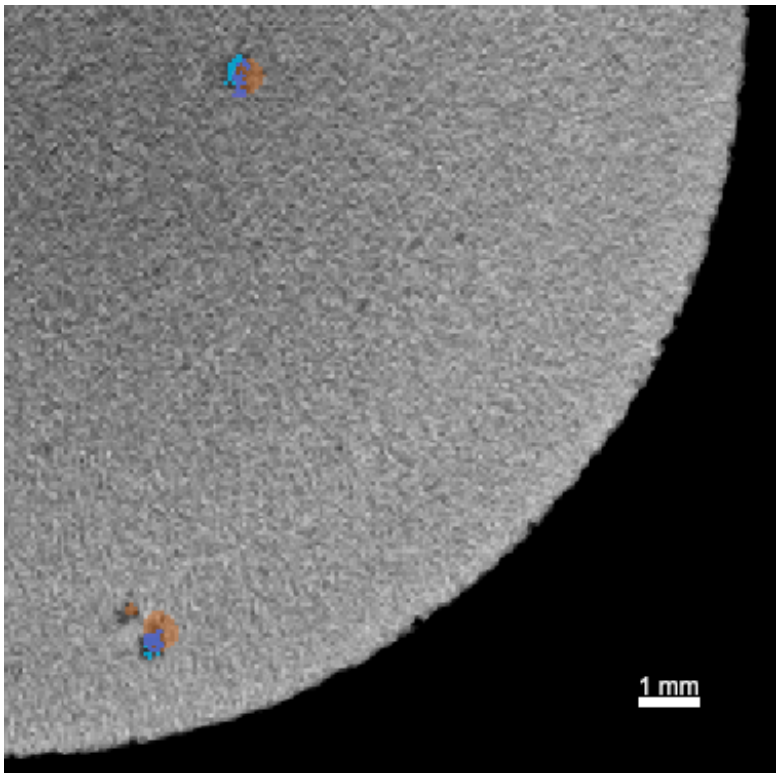


Figure 5.32 CT image of the sample B17 with the ground truth labels marked in blue and the neural network prediction for Epoch 8 shown in orange. The overlap of both label maps is visualized in purple. A good spatial correlation is visible.

The model shows the worst performance for specimen B23. In contrast to samples A11 and B17, the actual defects in B23 are mostly concentrated close to the surface of the specimen. Many anomalies are even connected to the outside, classifying them as open pores (see Figure 5.33 and Figure 5.35). This is in line with the parameter set used for creating the anomalies. B23 is printed using a reduced sky-writing delay. This altered parameter primarily influences regions in which the laser would normally perform a turn with sky-writing, i.e., the ends of a hatch. On the inside of the specimen, this is the case when the hatch line is reached. On the outside, this is the case at the border of the sample when the laser turns to continue lasering on the same sample.

Figure 5.35 shows an example of an open pore marked in blue. The five biggest undetected defects are all pores close to the surface with a connection to the outside. In contrast, six of the seven largest detected anomalies are located within the specimen without an opening to the surface. In total, eight of the 15 biggest anomalies are open pores, of which only one is detected. On the other hand, of the seven non-open pores, only one is not detected. This indicates a worse network performance for anomalies close to the surface. Additionally, the inference for B23 contains a lot of false positive labels, i.e., close to the surface of the sample. Figure 5.36 illustrates the condensed appearance of false positives in the outer region of the specimen. False positives can also be seen along the hatch lines. Figure 5.37 shows an example of accumulated false positives along the hatch line.

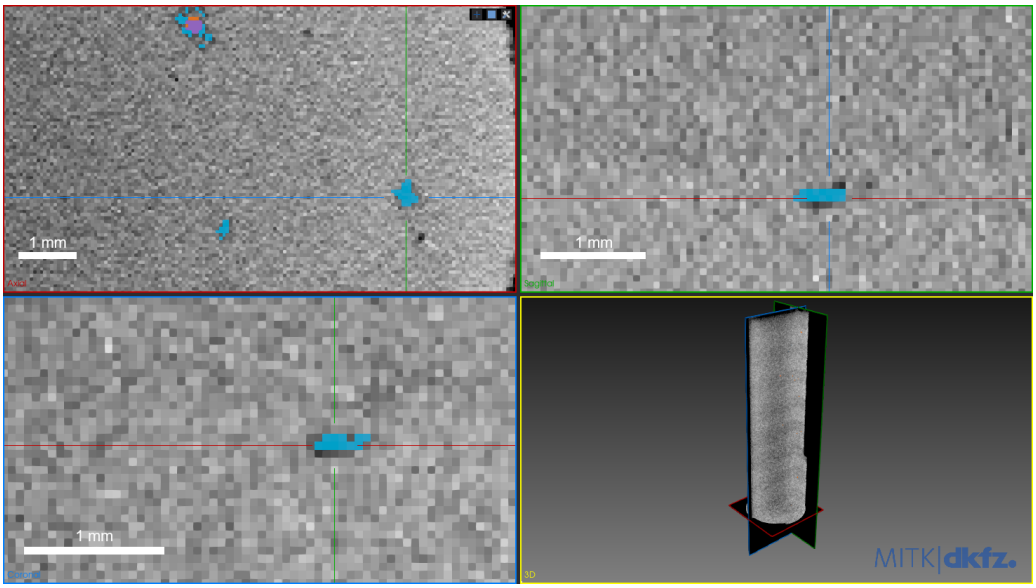


Figure 5.33 CT image of the BUD of the sample B17 with the ground truth labels marked in blue and the neural network prediction for Epoch 8 shown in orange. The overlap of both label maps is visualized in purple.

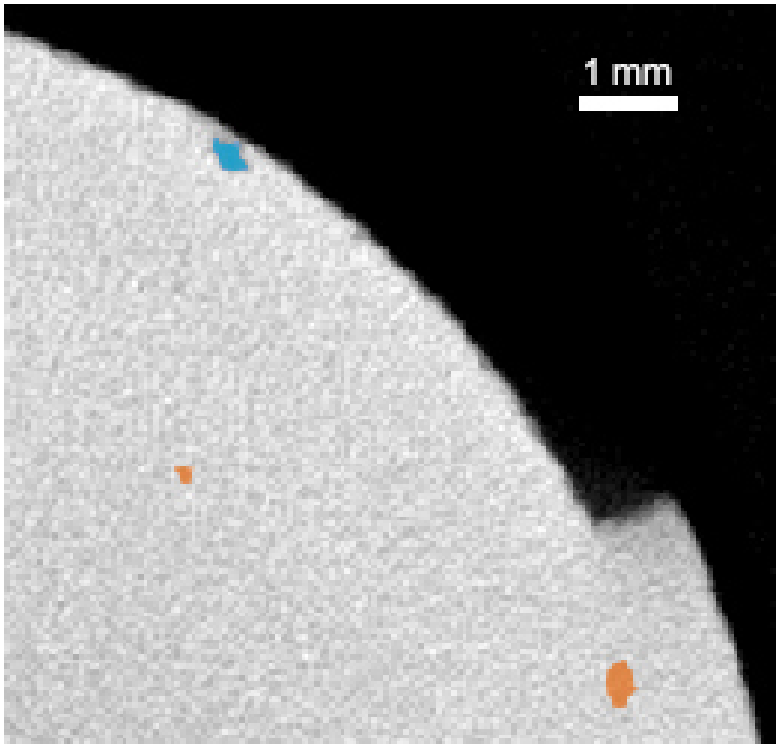


Figure 5.34 CT image of the sample B23 with the BUD marked in blue at the top right surface of the sample. The predictions of the 1. OM U-Net are marked in orange.

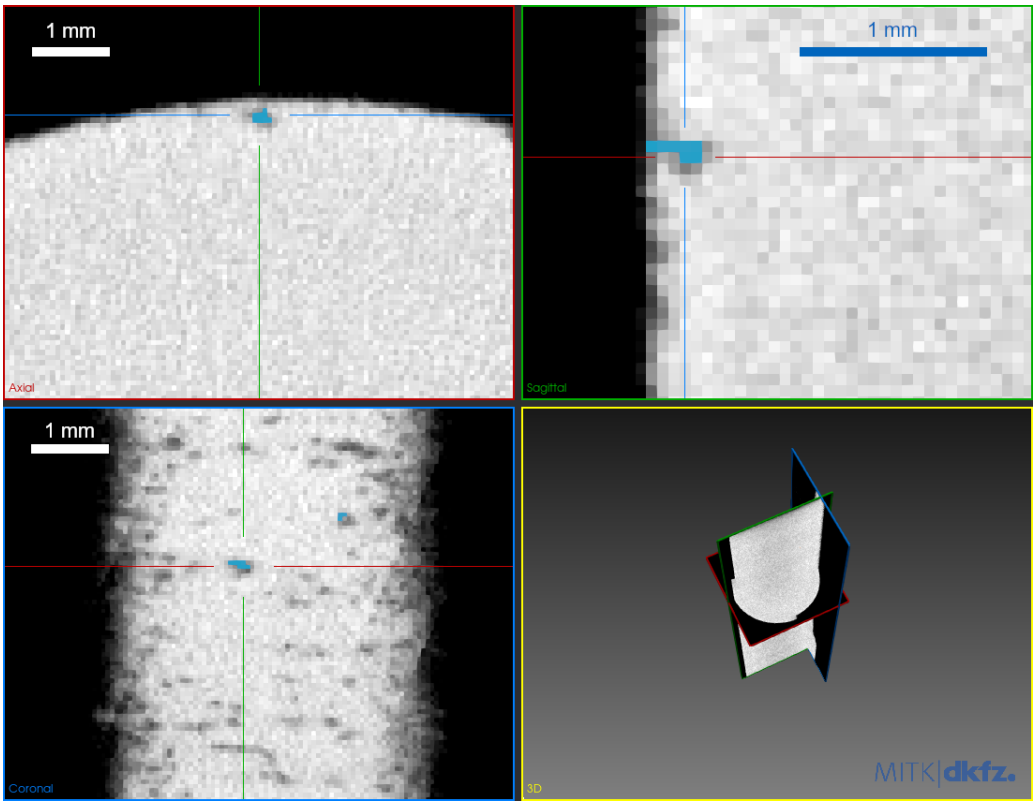


Figure 5.35 CT image of the sample B23 with an open pore marked in blue.

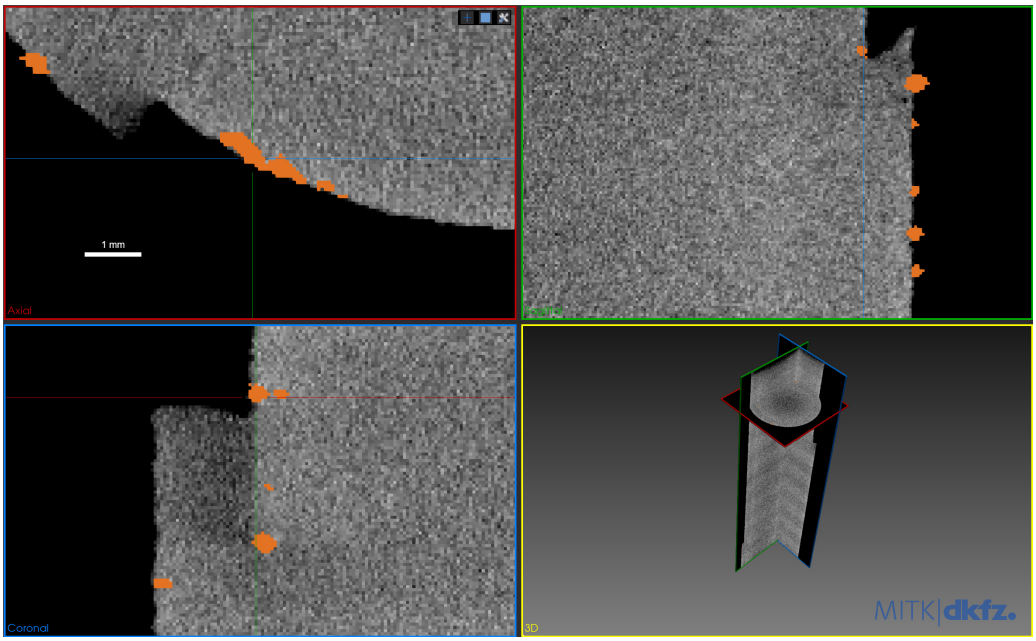


Figure 5.36 CT image of the sample B23 with false positive predictions close to the surface of the sample (Ground Truth: blue (not existent), Epoch 8: orange).

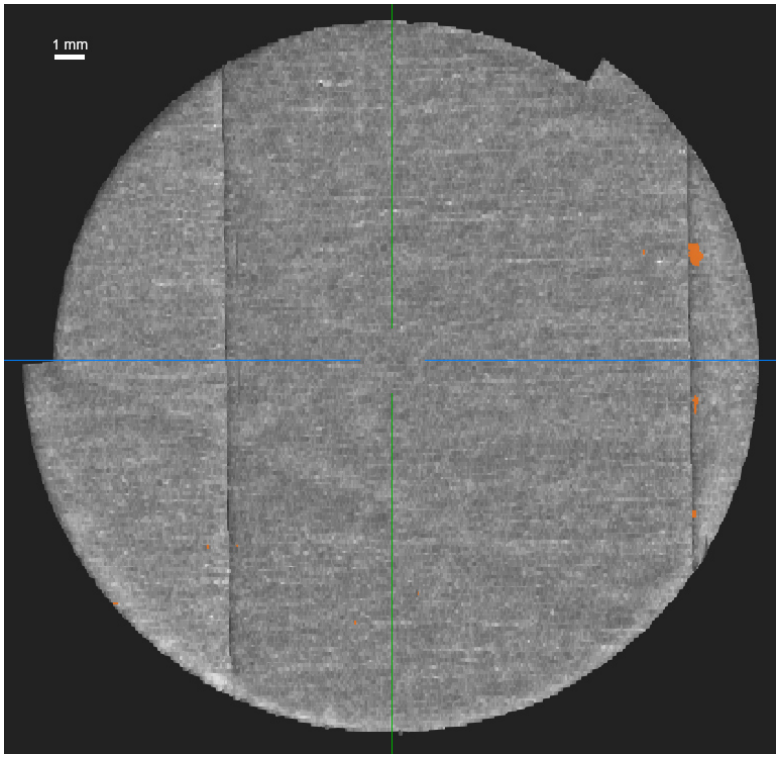


Figure 5.37 OM image of the sample B23 with false positive predictions along the right hatch line. The hatch lines are visible as two dark grey vertical lines within the specimen. (Ground Truth: blue (not existent), Epoch 8: orange).

There are different possible reasons for the lower detection performance at specific regions in the specimen. The data quality could be inferior at the border of the specimen. This seems unlikely as the monitoring data is gathered on-axis and should not deviate significantly for the border region of a sample. More likely, anomalies at the border of a sample might be underrepresented in the training dataset. Suppose the training dataset does not contain sufficient data points for a neural network to adapt to the specific feature of anomalies at the border. In that case, the neural network might perform inferiorly for this data point. In this case, the pores at the border of the sample might represent such a case. This would infer that a pore on the outside of a specimen represents a different feature for the neural network than a pore on the inside of the specimen. A possible mitigation is the collection of more data for this specific case. At the same time, the same behavior would be observable if the neural network is actively hindered from detecting pores on the outside of the sample. If the training data contains many anomalies open to the surface but not labeled as anomalies, the CNN would actively learn to disregard those regions as non-anomalies. In the inspection of AM parts by CT, a pore that is open to the outside is not necessarily classified as an anomaly. Such pores would generally be detected in a subsequent fluorescent penetrant testing step or removed by a surface treatment. Therefore, the qualified label maps contain many cases in which open pores on the outside are not labeled as anomalies. Part of this data is used to train the CT CNN, which, as a result, is trained to label such areas as anomaly-free as well. As the CT CNN generates the training data for the OM CNN, this behavior would be propagated to the final prediction behavior of the OM CNN. While this behavior is generally desired, it poses a challenge for the performance evaluation described in Section 4.3.4. A further explanation is the focus of the network on obvious features in the proximity of labels. In particular, for specimen B23, this could be the existence of hatch lines or the border of the sample. If the location of labels coincides with such obvious features for a relevant amount of time, the network might adapt to these features. Instead of focusing on more complex features that indicate pores, the network is then taught that the easily detectable hatch line often implies a defect location. As a result, the CNN would predict a significant number of false positives along the hatch lines. However, this assumption only holds true if the network either distinguishes globally between different printing parameter sets or is able to detect the altered skywriting behavior based on the monitoring data. Otherwise, it should also predict false

positives close to the surface for A11 and B17. As the OM data is normalized globally per specimen, the overall OM data should not show significant differences in intensity or distribution between the specimens. On the other hand, the adapted sky-writing delay should be easily detectable by the CNN as it alters the OM data at the end of the laser track as the laser is turned off later in the hatch turn. In the early epochs of the training, the network might adapt to this "obvious" feature in the data as it indicates a defect in a relevant number of times. Over the training process, the network is refined. It might be able to distinguish between the general pattern of the delayed sky-writing signal and features in the signal indicating an actual pore. To investigate this further, a later epoch, Epoch 20, is evaluated in the following.

Epoch 20 To analyze the development of the model performance over the training process, the inference results of Epoch 20 are evaluated in more detail as it shows the second lowest POD for B17 (1.39 mm) and an average POD for A11 (1.01 mm). For specimen A11, the BUD stays constant at 782 µm. The NUD_{400} decreases slightly from 20 to 17. Upon visual inspection, no significant changes are visible for the larger anomalies, which are detected in both epochs as illustrated in Figure 5.38.

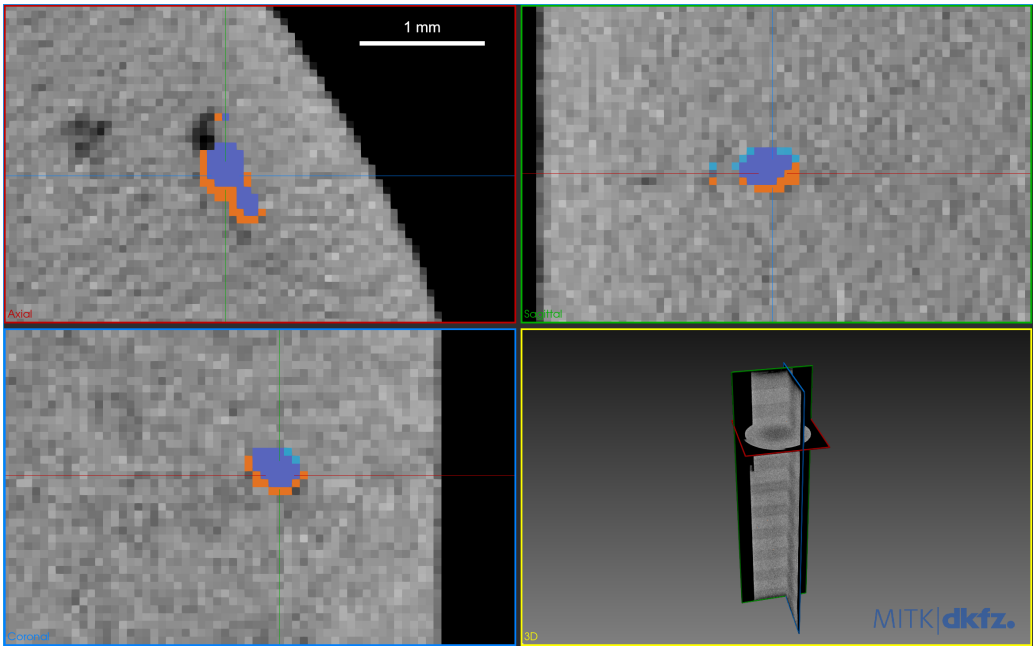


Figure 5.38 CT image of the sample A11 with the label map of Epoch 8 in orange, and the label map of Epoch 20 in blue. The intersection of both labelmaps is visible in purple. There are only slight changes visible between the two inference results.

An improvement is visible when inspecting the largest anomaly in A11, which is not detected by Epoch 8 but is found by Epoch 20. For Epoch 20, the model is able to detect and segment the anomaly with only minor deviations from the ground truth. In Epoch 8, the model did not mark any voxels in the surroundings as possible anomalies. On the other hand, the same behavior is also visible in Epoch 8, which detects anomalies not detected by Epoch 20. Hence, the network changes slightly over the training period but does not show significant visual changes for A11.

For specimen B17 the BUD increases from Epoch 8 (601 µm) to Epoch 20 (751 µm) with the NUD_{400} decreasing slightly from 29 to 27. The BUD for Epoch 20, which was previously detected in Epoch 8, is shown in Figure 5.39. The anomaly is an open pore to the surface. This highlights the previously described ambiguity for these defects and the challenge they pose to the CNN training.

The BUD and NUD_{400} for B23 do not change significantly either between the two epochs (BUD constant 556 µm, NUD_{400} from 12 to 15). However, the number of false positives larger than 400 µm decreases significantly from 642 (Epoch 8) to 287 (Epoch 20). This suggests that the CNN inference improves from Epoch 8 to Epoch 20. This is in contrast to the samples A11 and B17, for which no significant change in FP_{400} is noticeable. This supports the assumption of an impaired training process for pores in B23. As described above and shown in Figure 5.36, many false positives in B23 are located close to the surface.

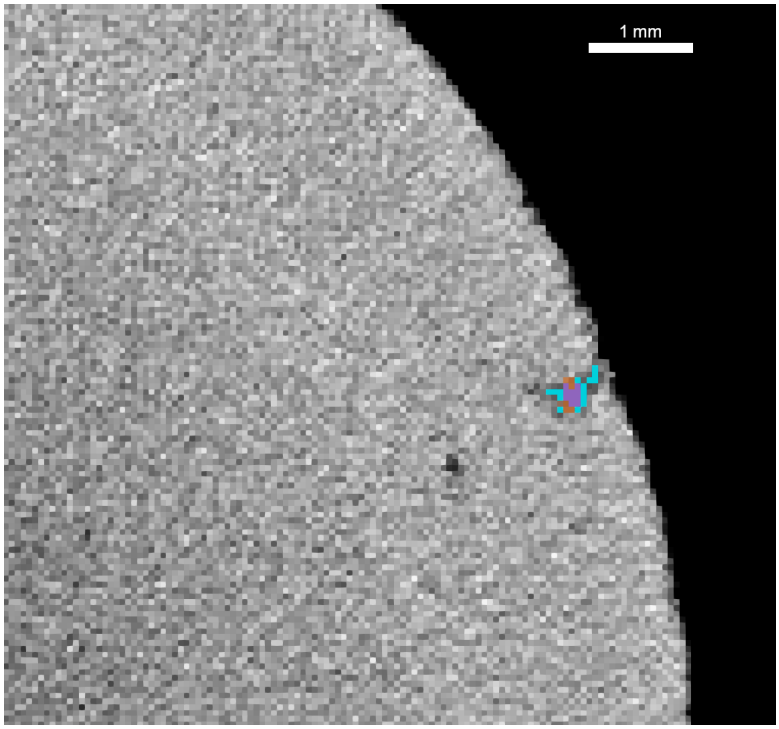


Figure 5.39 CT image of the sample B17 with the ground truth in blue and the label map of Epoch 8 in orange. The overlap of both label maps is visualized in purple. The anomaly is only found in Epoch 8 and not in Epoch 20.

For such pores, an ambiguity in the labeling is introduced together with a decreased representation in the training data and the concurrence with obvious features (e.g., hatch lines). All three effects might slow down or prevent the precise training of the model. Therefore, the model might require more epochs to train and hence show an improvement from Epoch 8 to Epoch 20. All three effects highlight the importance of data quality for training and testing and of a well-tuned training process that allows the CNN to adapt to difficult labels.

Despite the challenging cases described above, the overall feasibility of the proposed system could be shown. The 1. OM U-Net can detect defects in AM parts solely based on the online monitoring data. In the following approaches, the performance of the model and relevant limitations are explored.

5.2.2 2. OM U-Net: Performance Optimization

The second model trained on the OM dataset uses the same U-Net architecture as the 1. OM U-Net. The investigated approach tries to improve the performance of the model by extended data augmentation as described in Section 4.6.2.

Training results In the first step, analogous to Section 5.2.1, the training process of the CNN is evaluated using the training and validation loss as well as the validation Dice. The training and validation loss in Figure 5.40 and Figure 5.41 show a quick decrease within the first 4000 steps to around 0.35 and 0.1 respectively. Subsequently, both losses decrease significantly slower and start fluctuating around 0.3 and 0.09, respectively. The inverse behavior is visible for the validation Dice score (Figure 5.42). While this indicates a quick adaption of the model in the early training stages it is a slower decrease than for the 1. OM CNN. This indicates a longer adaptation process than for the 1. OM U-Net, which levels out within four epochs. This is to be expected as the stronger data augmentation creates more diverse training data, which in turn poses a more challenging task to the network. On the other hand, the more diverse and challenging training allows the model to adapt to more data points, which should improve the performance and, in particular, the robustness. This is evaluated based on the qualified label maps in the following.

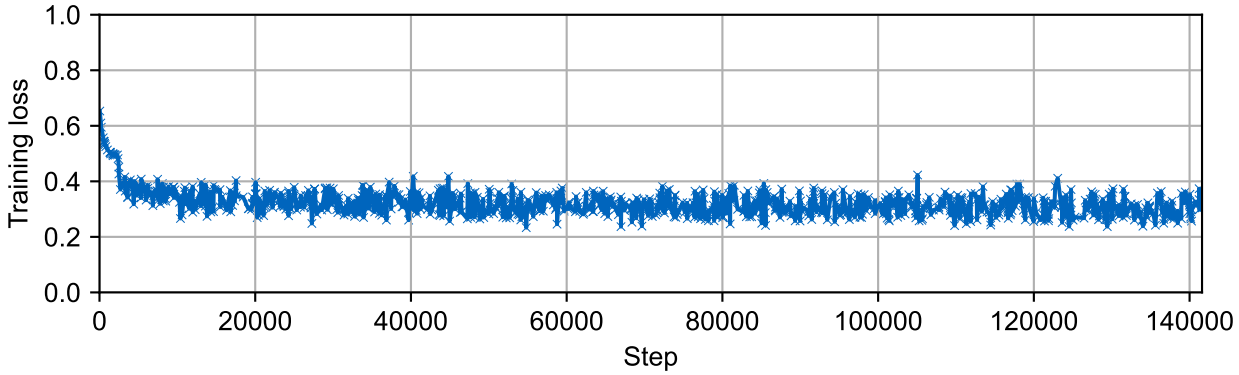


Figure 5.40 Training loss for the 2. OM CNN per training step.

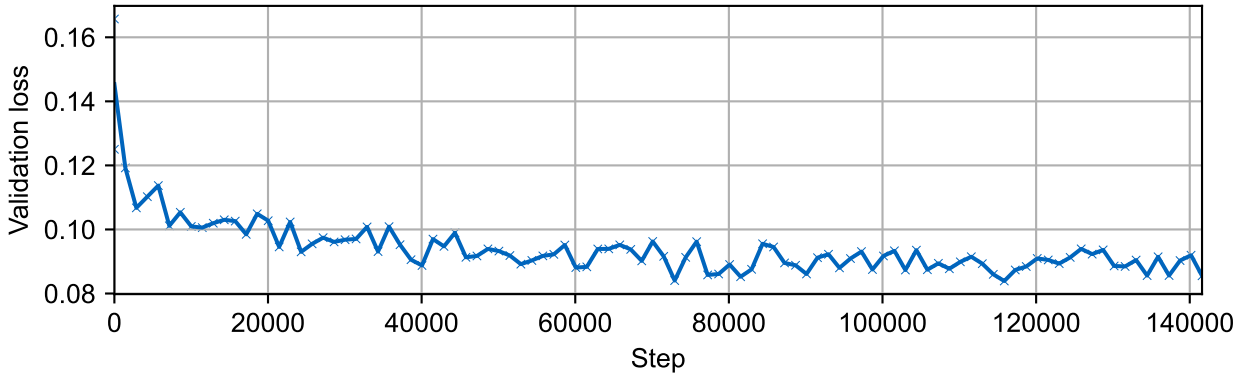


Figure 5.41 Validation loss for the 2. OM CNN per training step.

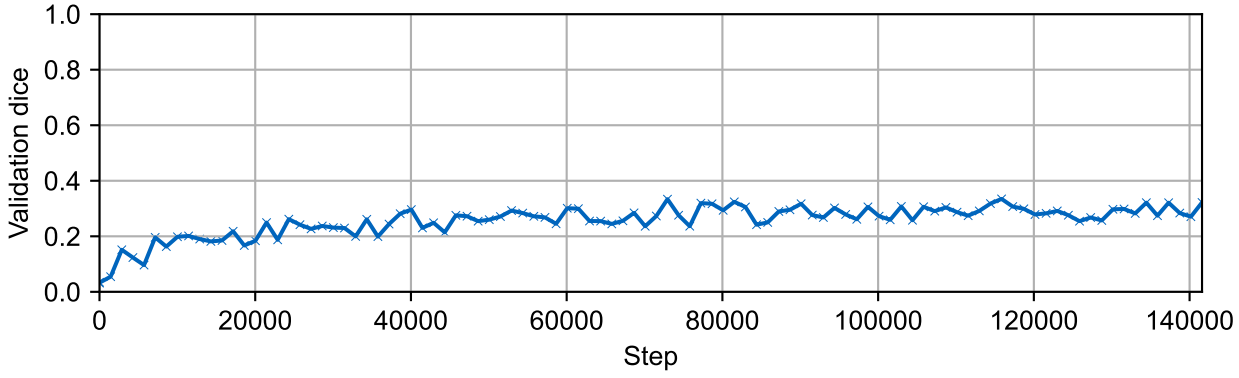


Figure 5.42 Validation Dice for the 2. OM CNN per training epoch.

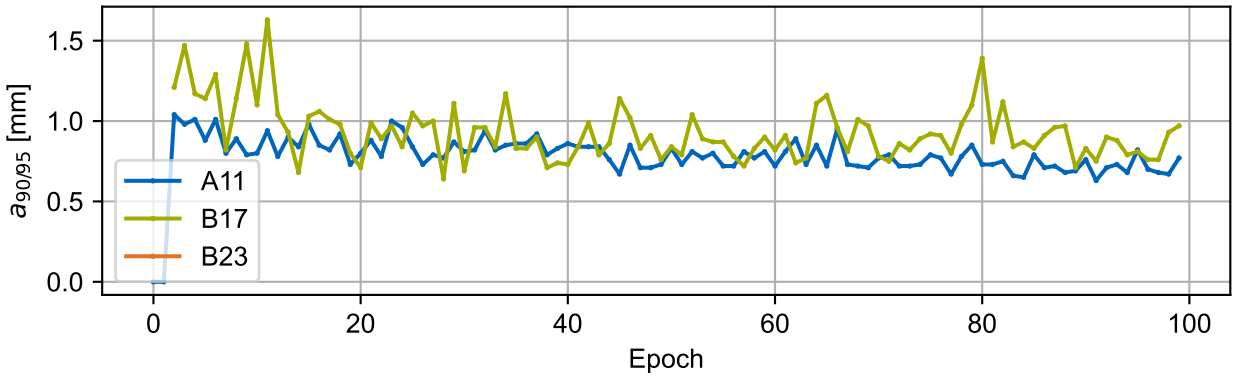


Figure 5.43 POD for the 2. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange, not calculable).

Pseudo-Test results The POD in Figure 5.43, the BUD in Figure 5.44 and the NUD_{400} in Figure 5.45 show significant differences between the specimens A11, B17 and B23. This is in line with the findings in Section 5.2.1, and as a result, the following interpretation is split on a per-specimen basis. The POD shows a qualitatively similar trend to Section 5.2.1. No POD can be calculated for any of the three specimens for the first two epochs. From Epoch 2 on, the POD can be constantly calculated for A11 and B17, while no POD can be determined for any epoch for B23. The POD for A11 reaches its minimum of 0.63 mm for Epoch 91. The lowest POD for B17 is 0.64 mm at Epoch 28. Overall, the POD for A11 shows a lower fluctuation than for B17.

The BUD for A11 drops quickly to 531 µm within the first epochs. It shows a low fluctuation in comparison to B17 and B23. The lowest BUD for A11 is 493 µm and is reached for multiple epochs. The BUD for B17 also decreases quickly but exhibits a stronger fluctuation for the first approximately 40 epochs. A lower variation is visible for the later epochs, with the lowest BUD of 439 µm reached at Epoch 70. B23 shows a jumping BUD over the entire training process. This indicates a strongly varying network performance for specimen B23, which aligns with the findings in Section 5.2.1. The lowest BUD is 487 µm at Epoch 83. This epoch also represents the lowest overall BUD for all three specimens with 531 µm.

The same trend for the different samples is visible for the NUD_{400}. The specimens A11 and B17 show a quick drop to around 10% while the NUD_{400} for B23 stays at around 50 % with a strong fluctuation. For A11 and B17, this means that around 90 % of all relevant defects are detected. Considering the ambiguous label map for some defects, this can be regarded as a promising result.

The FP_{400} is additionally considered to select individual epochs for further investigation. Figure 5.46 plots logarithmically the FP_{400} for A11, B17 and B23. With around 100 false positives, the model produces significantly less false positives on the specimen A11. For B17, the FP_{400} stays at around 500 while the false positive count for B23 fluctuates strongly between approximately 50 and 1500. When comparing the false positive count with the NUD_{400}, the trade-off between the two metrics is visible as for the 1. OM U-Net. For all three specimens, epochs with a low FP_{400} record a high NUD_{400} and vice versa.

Interpretation In the following, Epoch 91 is analyzed in detail as it produces the lowest POD (0.63 mm), BUD (493 µm) and NUD_{400} (2%) for A11. In addition, a relatively low FP_{400} of 190 is achieved compared to other epochs with a low POD. The quantitative analysis indicates a well-performing network that can detect 98% of all relevant defects in A11. Hence, it is regarded as the best epoch with respect to its overall performance and the trade-off described above.

Figure 5.47 shows the BUD for Epoch 91 on specimen **A11**. The defect matches the pattern described in Section 4.3.4 and Section 5.2.1 for ambiguous labels. The same holds true for the second largest undetected defect. These are the only two undetected defects larger than 400 µm.

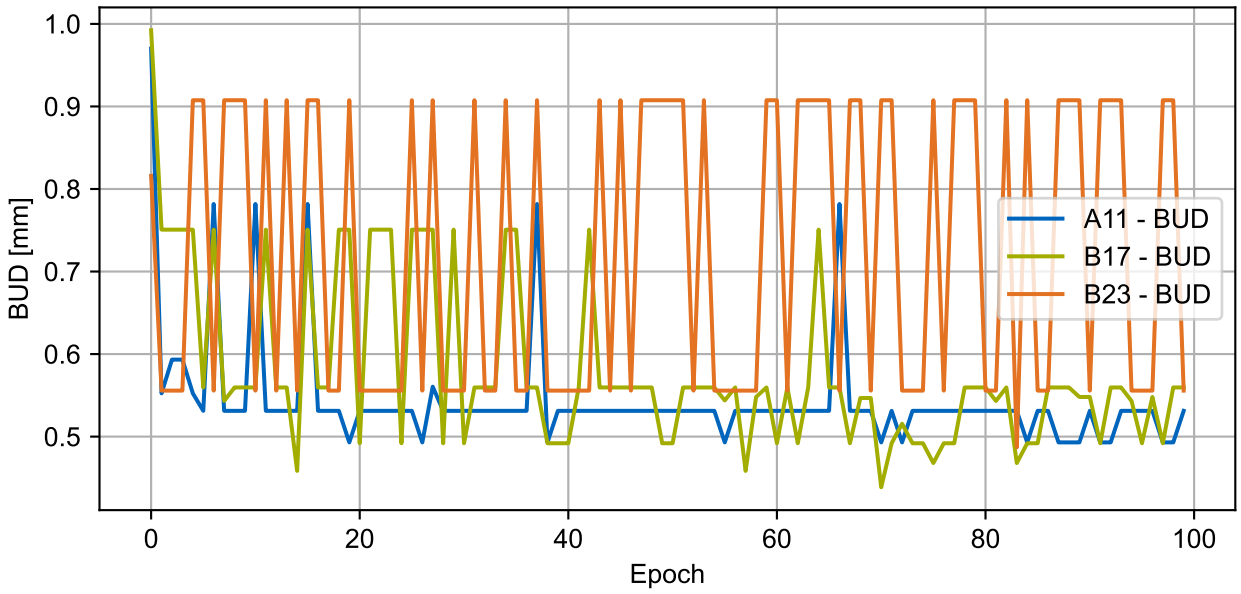


Figure 5.44 Biggest undetected defect for the 2. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange).

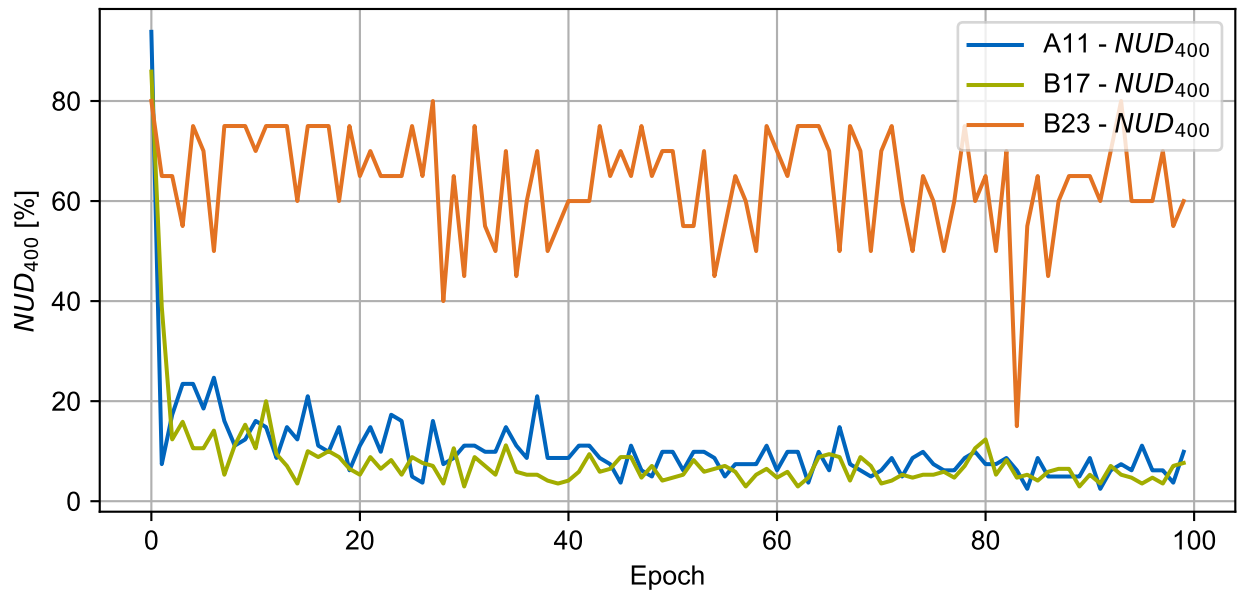


Figure 5.45 Percentage of undetected defects larger than 400 µm for the 2. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange).

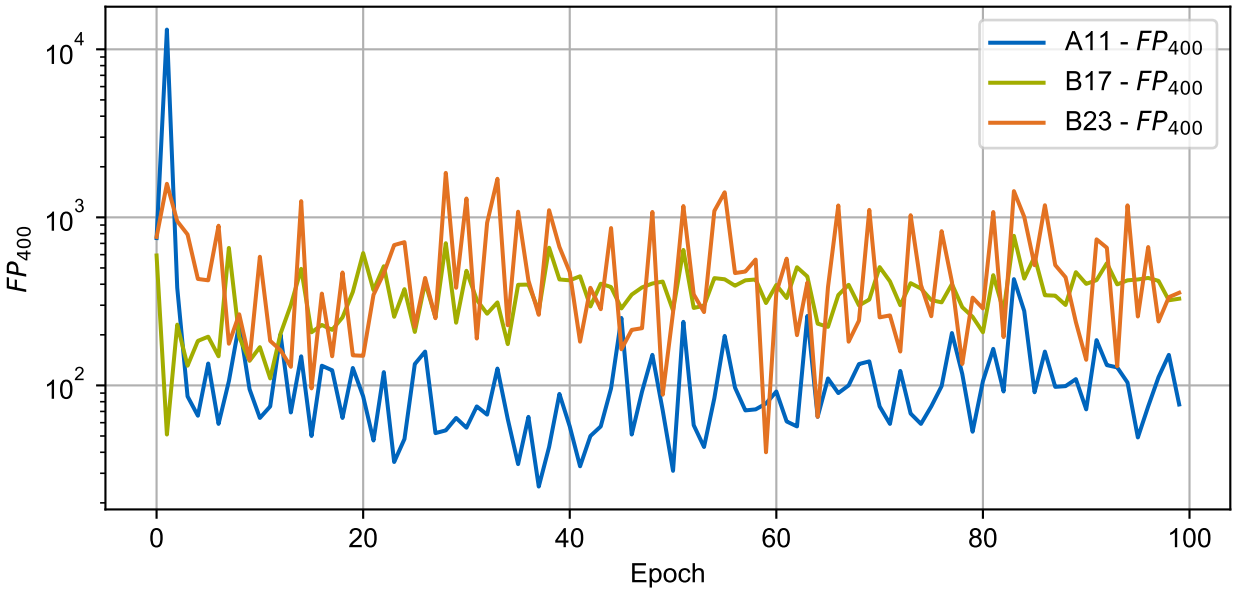


Figure 5.46 Number of false positive detections larger than 400 µm plotted logarithmically for the 2. OM CNN per epoch on samples A11 (blue), B17 (green), and B23 (orange).

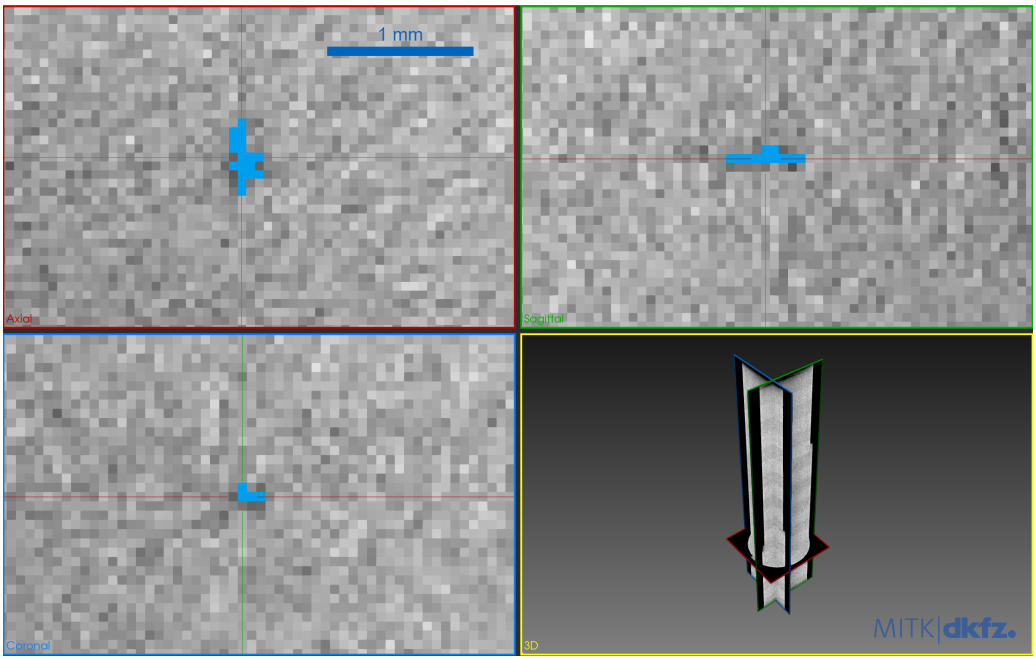
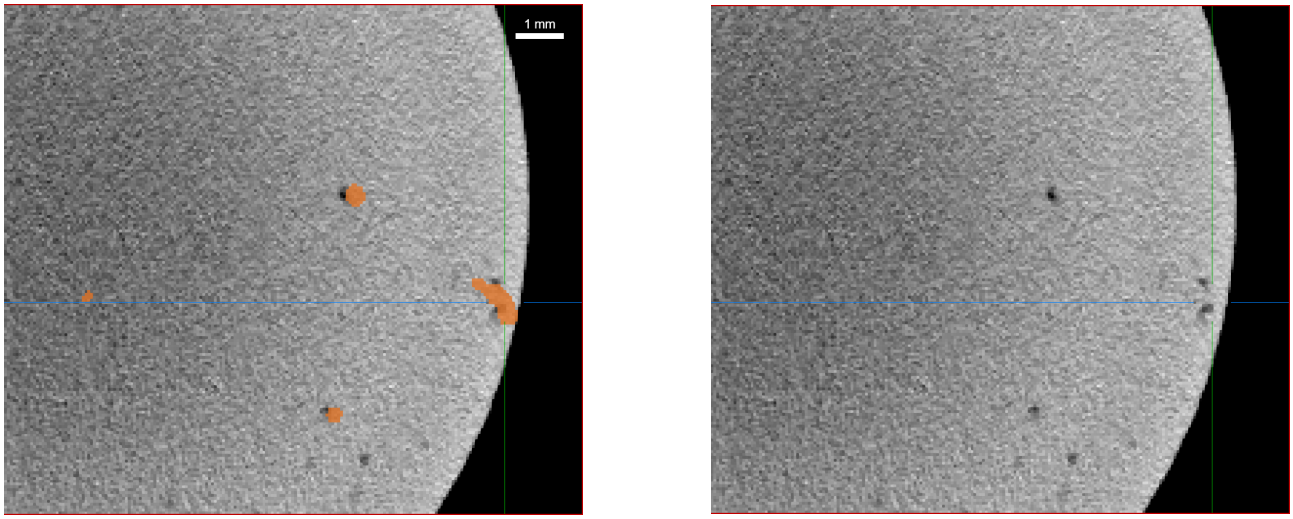


Figure 5.47 The biggest undetected defect for Epoch 91 of the 2. OM U-Net on A11. The GT is shown in blue.



(a) With the prediction marked in orange.

(b) Without the label map overlay.

Figure 5.48 The largest false positive for the 2. OM U-Net for Epoch 91 on A11. A possible anomaly is visible but is not labeled according to the ground truth.

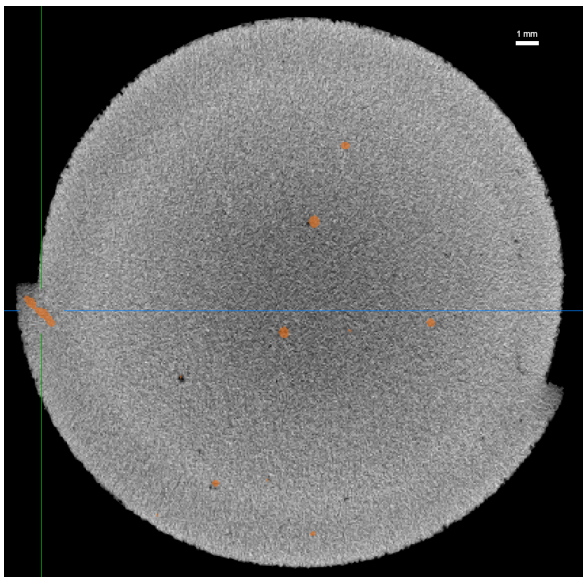
When investigating the largest false positives produced on A11, the same ambiguity becomes visible. Figure 5.48 plots the prediction of the 2. OM U-Net in orange next to the unlabelled CT in the same position. The labeled region does exhibit signs of an anomaly in the CT but is not classified as a defect according to the qualified ground truth. The distinction of such border cases is challenging for the network, particularly when considering the inherent ambiguity of the training and test label maps. Future steps to refine this behavior are discussed in Section 6.3.

The POD for **B17** at Epoch 91 is $0.75\,\mathrm{mm}$ which lies $0.11\,\mathrm{mm}$ above the minimum POD of $0.64\,\mathrm{mm}$. The BUD is $492\,\mu\mathrm{m}$, which is slightly above the minimum of $439\,\mu\mathrm{m}$ at Epoch 70. The NUD_{400} is 3 % while the FP_{400} is 424.

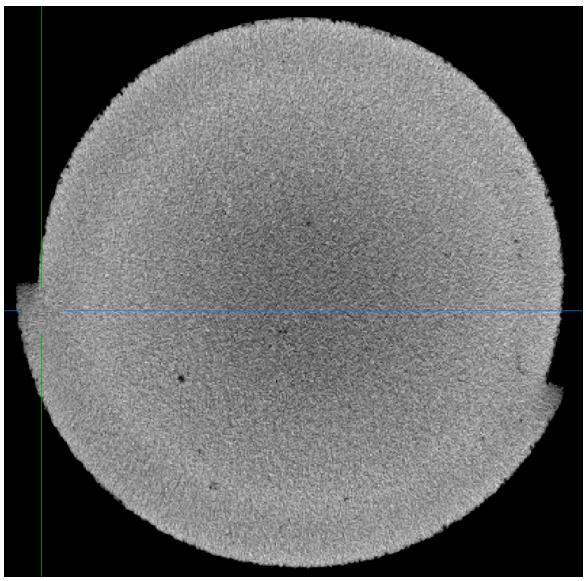
Based on visual inspection of the inference results, the high false positive count can be partially attributed to the ambiguous label map. While the largest false positive is a clear false prediction close to the surface of the specimen, the next biggest false positives represent difficult border cases with ambiguous anomalies visible in the close vicinity. This is illustrated in Figure 5.49 and Figure 5.50. Figure 5.49 contains the largest false positive on the left side of the specimen. Additionally, a number of ambiguous false positives with possible anomalies in their vicinity are visible in the center of the sample and in Figure 5.50. The same influence of ambiguity holds true for the false negatives. The five largest undetected defects of the seven undetected defects larger than $400\,\mu\mathrm{m}$ represent ambiguous anomalies (i.e., open to the surface or less than three slices in z-direction).

The model at Epoch 91 is not well suited to analyze the specimen **B23**. It does not detect the largest defect in the sample and only detects about 40% of all defects larger than $400\,\mu\mathrm{m}$. Furthermore, there are around 750 false positive predictions in the inference. This is also supported by a visual inspection of the inference label map. It shows a high concentration of false positives along the surface of the specimen with no corresponding anomalies in their vicinity. This indicates the same inferior performance compared to A11 and B17 as for the 1. OM U-Net, and supports the assumptions made in Section 5.2.1.

Overall, the 2. OM U-Net can be regarded as an improvement to the 1. OM U-Net. The POD, BUD, and NUD_{400} are reduced significantly for A11 and B17. Furthermore, in comparison to the 1. OM U-Net, the number of systematic false positives for B23 along hatch lines has strongly decreased, even if the overall number of false positives has increased. The overall increase is partially attributed to the ambiguous ground truth label map. This is an improvement with respect to Section 5.2.1 as it indicates that the model is now able to better distinguish between the defect signature and irrelevant process signature (such as the hatch lines). Further steps for improvement are discussed in Chapter 7.

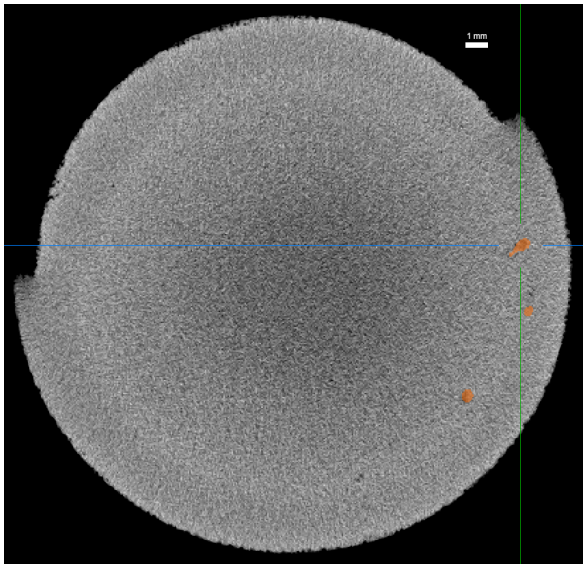


(a) With the prediction marked in orange.

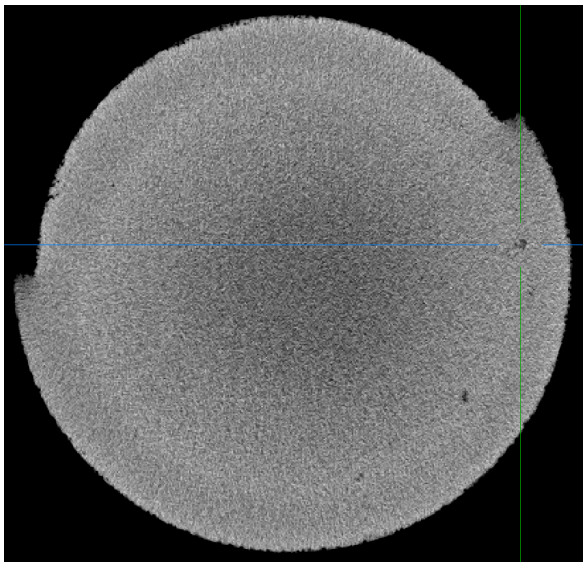


(b) Without the label map overlay.

Figure 5.49 The largest false positive for the 2. OM U-Net for Epoch 91 on B17. The largest false positive in the left part of the specimen does not show any visible anomalies in its vicinity. In contrast, the smaller false positives in the center of the specimens correspond to possible anomalies that are not labeled as defects according to the qualified ground truth.



(a) With the prediction marked in orange.



(b) Without the label map overlay.

Figure 5.50 The second largest false positive for the 2. OM U-Net for Epoch 91 on B17. The false positives in the center of the specimens correspond to possible anomalies that are not labeled as defects according to the qualified ground truth.

Considering all validation and test metrics, data augmentation for the use of CNNs in online monitoring shows the potential for performance improvements. By creating more diverse training data, data augmentation challenges the network more strongly during training. Hence, it requires more epochs to fully adapt to the training data. At the same time, the more diverse training data allows for a more robust network that is able to produce more accurate inference results for a larger variety of samples. This is reflected in the POD improvement for A11 from $0.78\,\mathrm{mm}$ for the 1. OM U-Net to $0.63\,\mathrm{mm}$ for the 2. OM U-Net and the corresponding visual inspections. For B17, the POD improvement is even greater, with the best POD for 1. OM U-Net being $1.32\,\mathrm{mm}$ in comparison to $0.64\,\mathrm{mm}$ for the 2. OM U-Net. Therefore, the CNN performance shows great potential with regard to the use-case-defined metrics. In conclusion, the 2. OM U-Net underlines the feasibility of the presented approach to detect defects based on a CNN analysis of the monitoring data. The influence of the different specimens on the CNN performance is further investigated in Section 5.2.4.

5.2.3 3. OM U-Net: Individual Channels

The third training approach investigates the relevance of the individual channels. As described in Section 4.6.3, it consists of six individual models, which are each trained on a single channel (TED^{max}, TED^{min}, TEP_{high}^{max}, TEP_{high}^{min}, TEP_{low}^{max}, and TEP_{low}^{min}). In the following, the individual training results are presented, and the performance of the individual channels is compared.

Training results Figure 5.51 shows the training loss of all six models. The curve is averaged per epoch to allow for a better visualization of the overall trend. The trainings for TEP_{high}^{max} and TED^{max} terminated early at Epoch 94 and Epoch 91, respectively, due to a server error. For TED^{max}, the training is continued until Epoch 99, but no significant improvements are visible. Hence, a rerun of the experiment does not seem necessary as no significant change in training and validation loss is observable for the later training stage.

The curves clearly show different training behaviors for the two models trained on the TED^{min} and TED^{max} channels. For the first training steps, the training loss decreases similarly for all six CNNs, but from around Epoch 4, the training loss for the two TED channels levels out at around 0.5. The other four models show a significantly lower training loss of around 0.35. All training losses fluctuate around the respective value, which is visualized by the standard deviation per epoch in lighter colors in Figure 5.51.

The validation loss curve in Figure 5.52 shows a similar trend to the training loss. The models trained on the TED metrics show a worse validation loss than TEP_{high} and TEP_{low}. In contrast to the training loss, the validation loss shows an additional differentiation between TEP_{low} and TEP_{high}. TEP_{low}^{min} shows a higher validation loss than TEP_{high} and TEP_{low}^{max}. This behavior is also visible for the validation Dice in Figure 5.53.

All three curves point towards a significantly worse network performance when only trained on either TED channel. The difference between TEP_{low} and TEP_{high} is less pronounced and is investigated further in the following, together with the actual use-case-specific performance of the models.

Pseudo-Test results Analyzing each of the six models based on the use-case-specific metrics allows for a direct comparison of the individual performances. In the following, the BUD and NUD_{400} are investigated per specimen. Important to note in this context is the sole training of the models on Buildjob A. Hence, a better performance on specimens from this buildjob is to be expected. The following evaluation focuses on the comparison of the different channels. The overall performance is of secondary importance as the main objective is the analysis of the influence of different channels.

For specimen A11, the BUD lies between $970\,\mu\mathrm{m}$ (which is the overall biggest defect in the sample) and $782\,\mu\mathrm{m}$ (see Figure 5.54). From an early epoch on, the CNNs trained on channels TEP_{high}^{min}, TEP_{high}^{max} and TEP_{low}^{min} show a constant BUD of $782\,\mu\mathrm{m}$. The model trained on TEP_{low}^{max} requires more epochs before constantly detecting this BUD as well. The models for $TED^{max}$and TED^{min} are not able to detect the

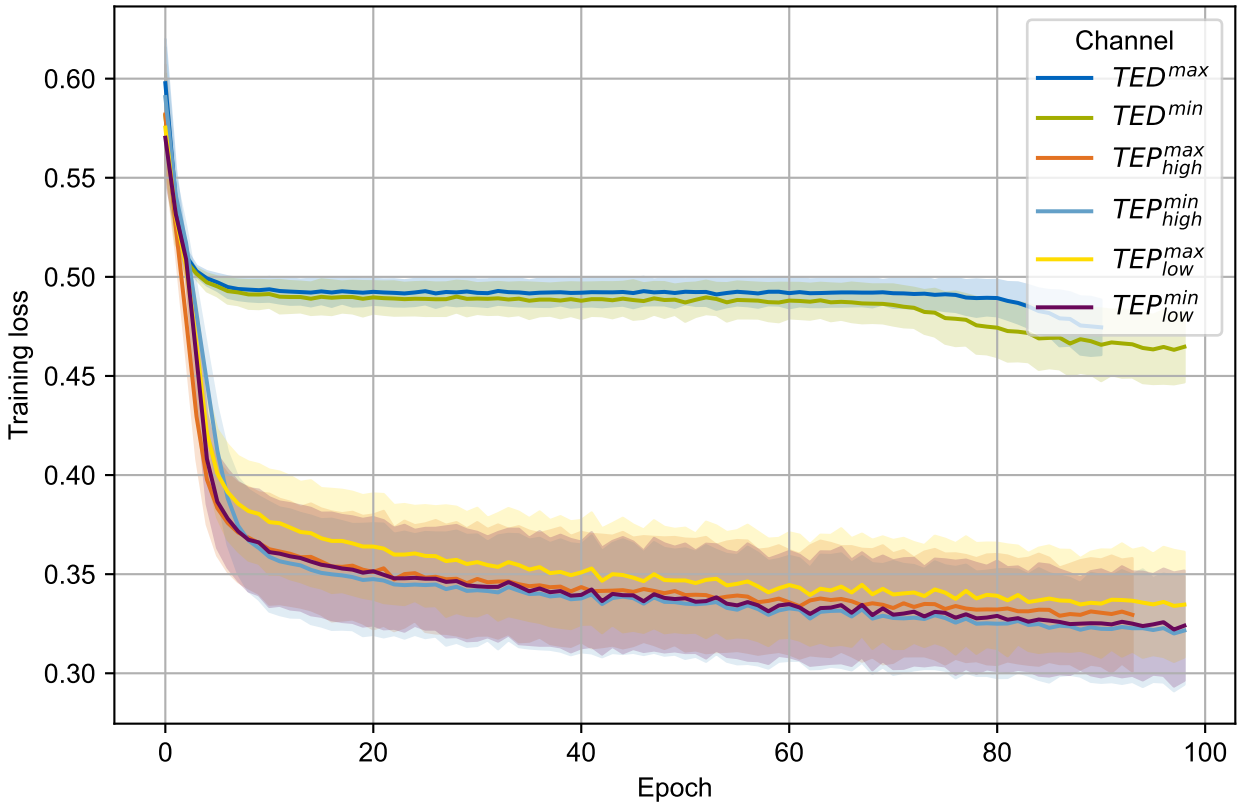


Figure 5.51 The training loss for each of the six CNNs trained on individual channels. The loss values are averaged per epoch, and the standard deviation per epoch is visualized by the lighter color bar.

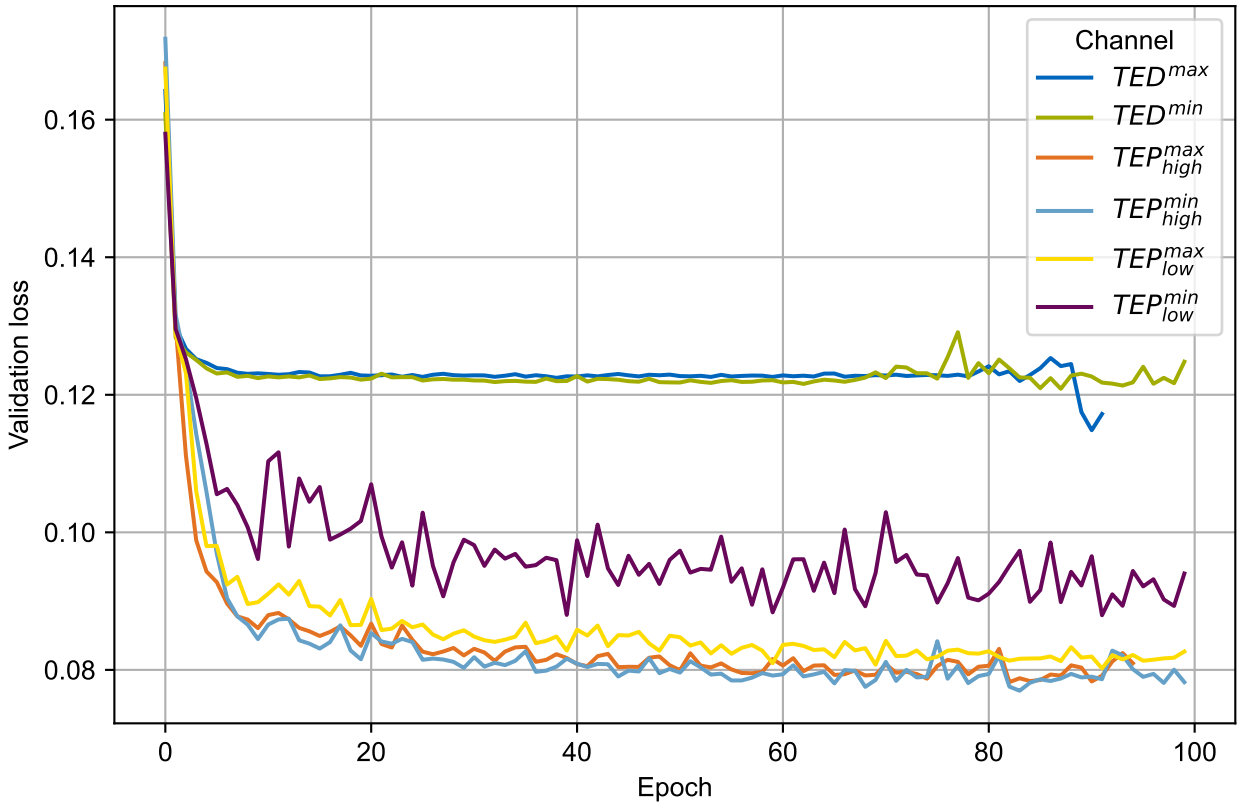


Figure 5.52 The validation loss for each of six CNNs trained on individual channels.

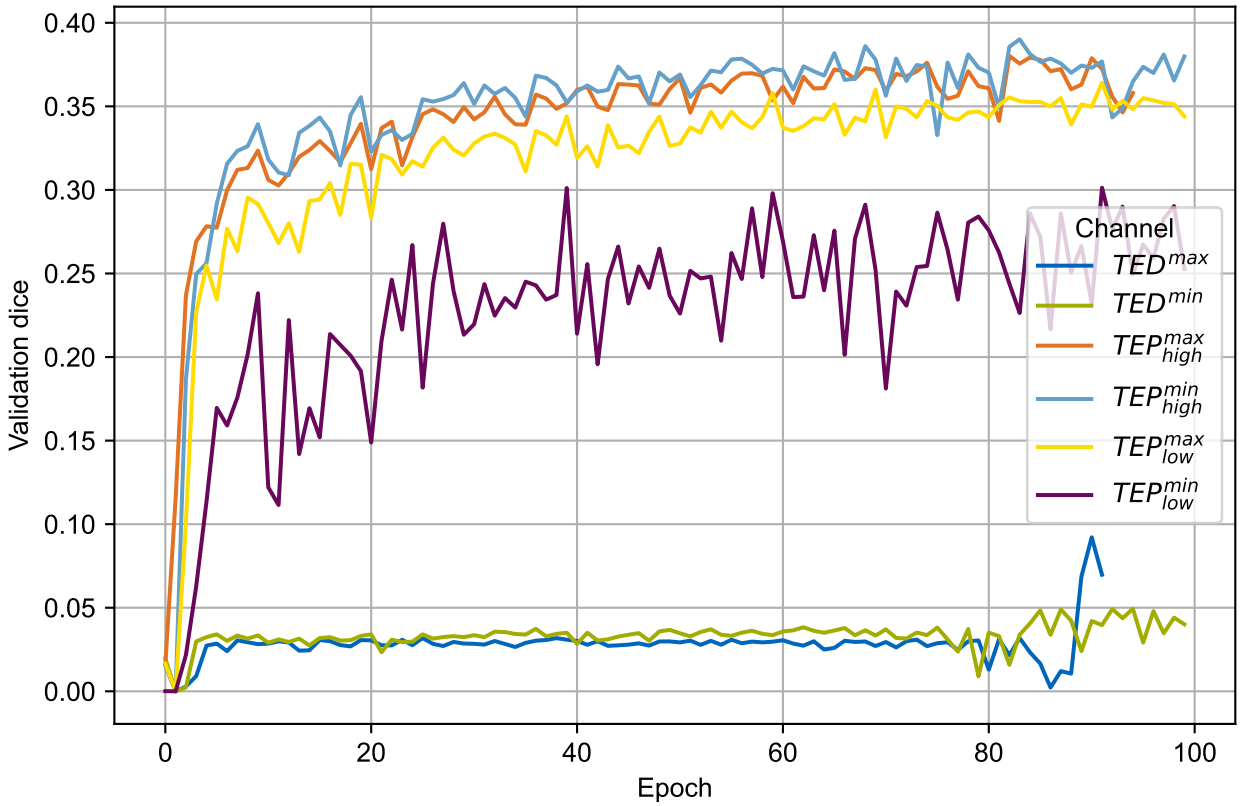


Figure 5.53 The validation Dice for each of six CNNs trained on individual channels.

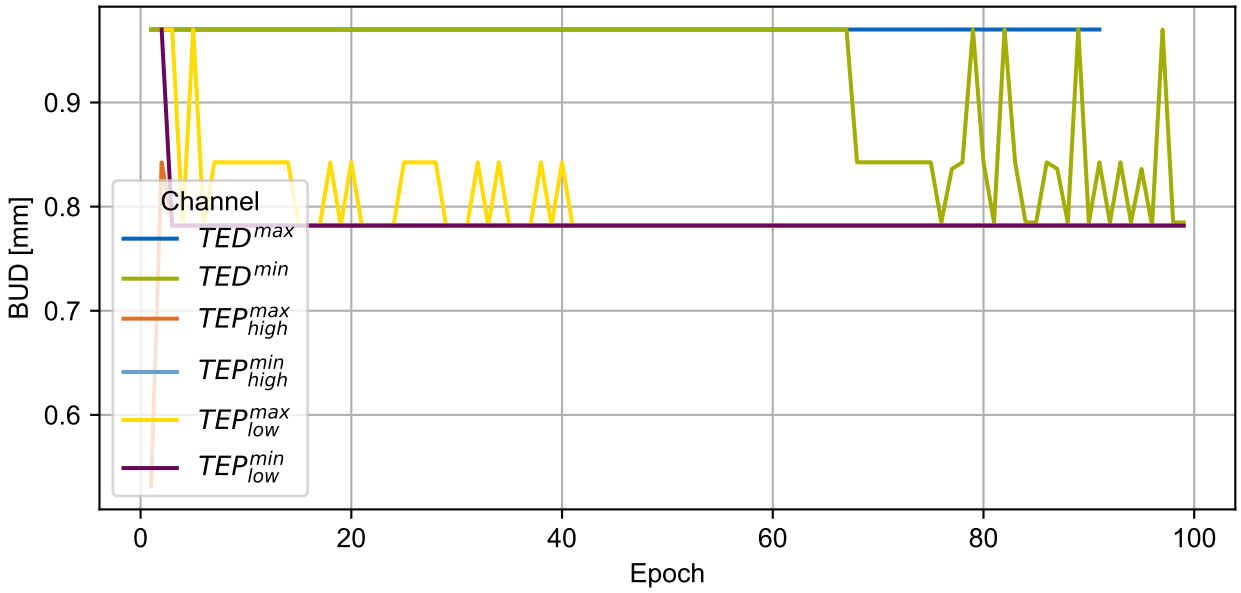


Figure 5.54 The biggest undetected defect in specimen A11 per epoch for all six CNNs trained on individual channels.

largest anomaly in A11 up to Epoch 67. For the later epochs, TED^{min} shows lower BUDs but only in combination with a very high number of false positives, as seen in Figure 5.55.

Figure 5.55 shows the number of false positives larger than 400 µm and confirms the differences between the channels already observed based on the BUD. The CNNs based on the TED^{max} and TED^{min} channels produce significantly more false positives in sample A11. The other four channels do not show substantial differences in the number of false positives with respect to each other.

Both observations imply a more robust and better-performing network for the TEP_{low} and TEP_{high} channels. The same applies to the NUD_{400} plotted in Figure 5.56.

The metrics for specimen B17 are plotted in Figure 5.57, Figure 5.58 and Figure 5.59. A similar inferior behavior is visible for the channels TED^{max} and TED^{min}. TED^{min} shows a stronger fluctuation in later epochs. However, the overall performance is still regarded as inferior, as an improvement in NUD_{400} corresponds with a degradation in the false positive count and vice versa. The channels TEP^{min}_{high}, TEP^{max}_{high}, and TEP^{max}_{low} resemble the same qualitative trend as for A11 for the BUD and the NUD_{400}. In contrast, the BUD and NUD_{400} for TEP^{min}_{low} show an inferior performance compared to TEP_{high} or TEP^{max}_{low}. On the other hand, the number of false positives larger than 400 µm for the model trained on TEP^{min}_{low} is significantly lower, indicating a better performance concerning false positive detections.

The pseudo-testing on specimen B23 shows an unclear picture with many false positives and high fluctuation for the BUD and NUD_{400} (see Figure 5.60, Figure 5.61 and Figure 5.62). As for the other samples, the model trained on TED (min or max) shows the worst performance. However, the other four channels also result in models with many false positives in combination with a high BUD and NUD_{400} compared to A11 and B17. Upon visual inspection of the TEP^{min}_{high} results (Epoch 88) on B23 the inference shows meaningful predictions but with a low quantitative performance. This is illustrated in Figure 5.63.

Interpretation Overall, the models trained on the different channels show significant differences in their performance. The TED^{min} and TED^{max} channels result in a consistently worse-performing network. This indicates that the TED channel provides less relevant information than the TEP_{low} and TEP_{high} channels, even though the TED channel captures the meltpool radiation in a broad spectrum, including the narrow bandwidth observed by TEP_{low} and TEP_{high}. Hence, it also contains the information included in TEP_{low} and TEP_{high}. Therefore, it is assumed that the inferior information extraction is due to an inferior

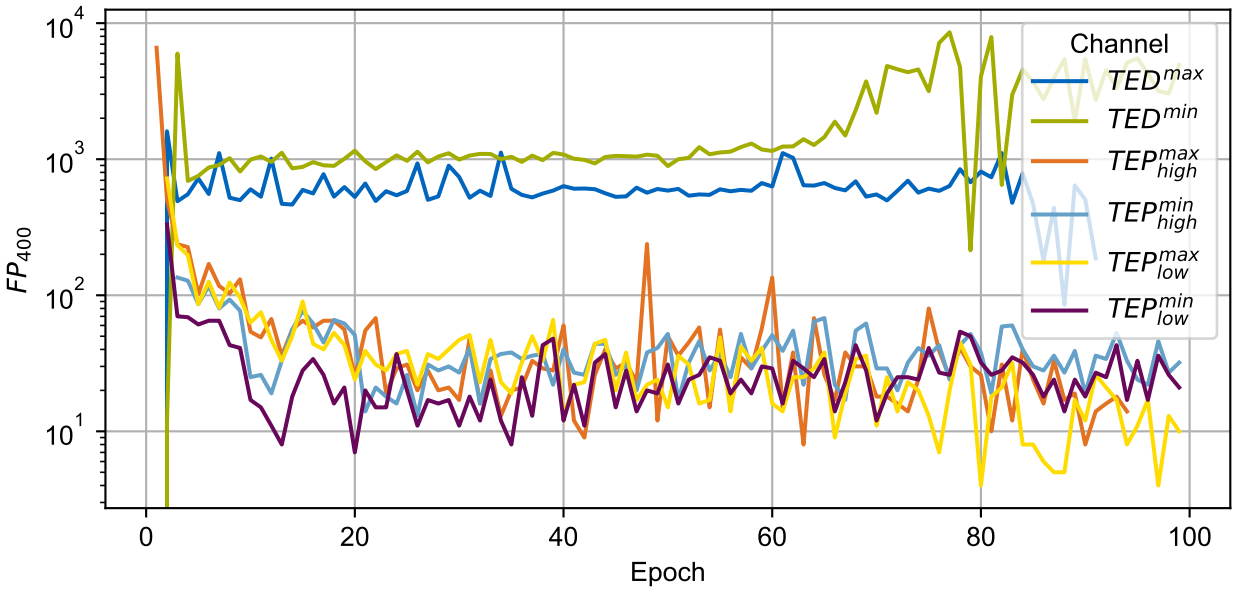


Figure 5.55 Number of false positives larger than 400 µm in specimen A11 per epoch for all six CNNs trained on individual channels (plotted logarithmically).

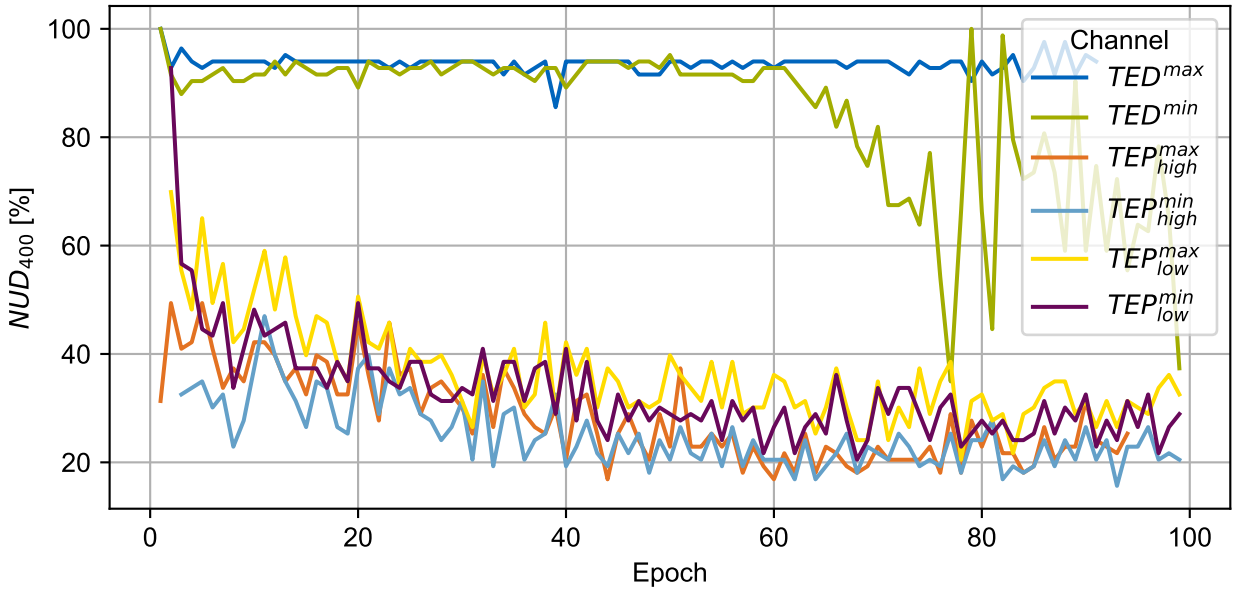


Figure 5.56 Number of false negatives larger than 400 µm (NUD_{400}) in specimen A11 per epoch for all six CNNs trained on individual channels.

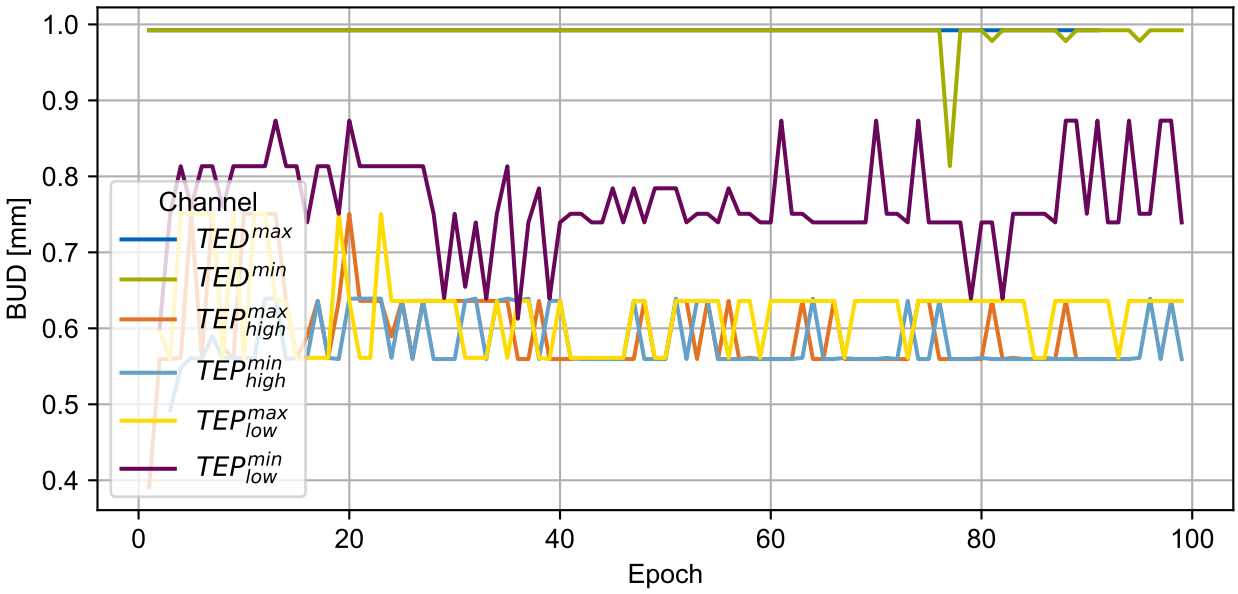


Figure 5.57 The biggest undetected defect in specimen B17 per epoch for all six CNNs trained on individual channels.

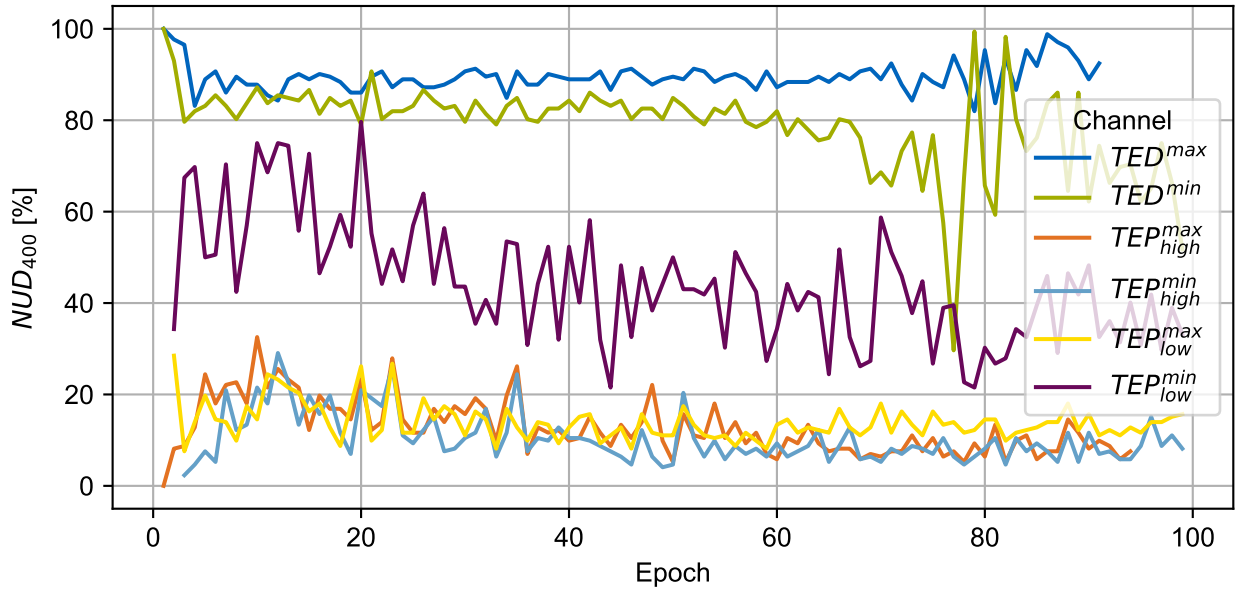


Figure 5.58 Number of false negatives larger than 400 µm (NUD_{400}) in specimen B17 per epoch for all six CNNs trained on individual channels.

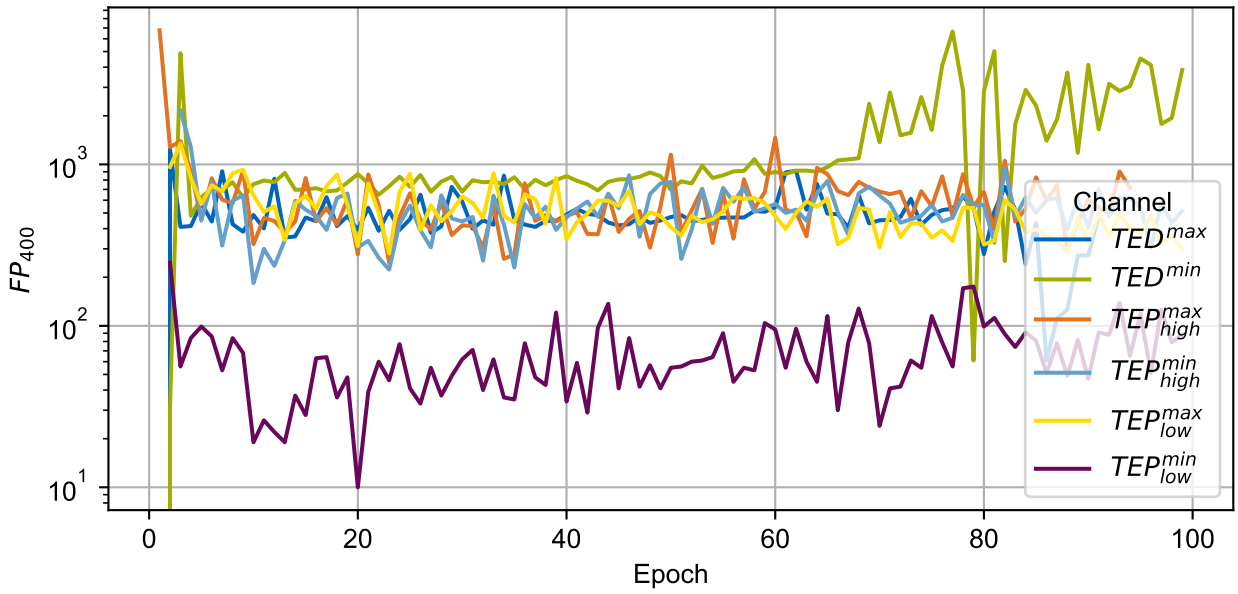


Figure 5.59 Number of false positives larger than 400 µm in specimen B17 per epoch for all six CNNs trained on individual channels (plotted logarithmically).

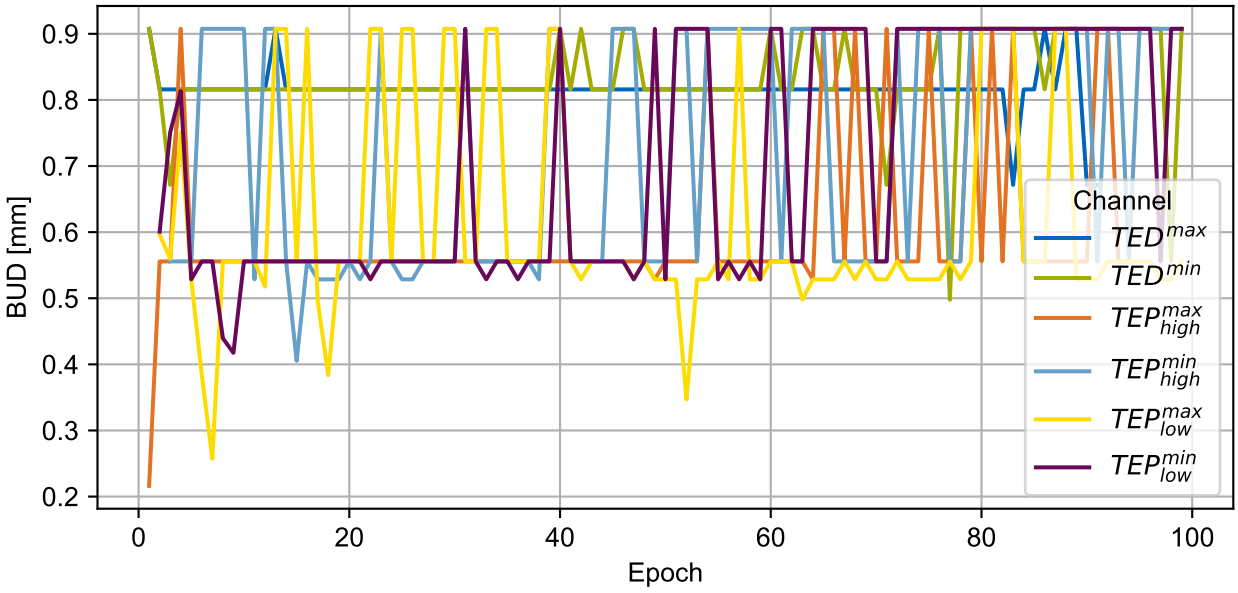


Figure 5.60 The biggest undetected defect in specimen B23 per epoch for all six CNNs trained on individual channels.

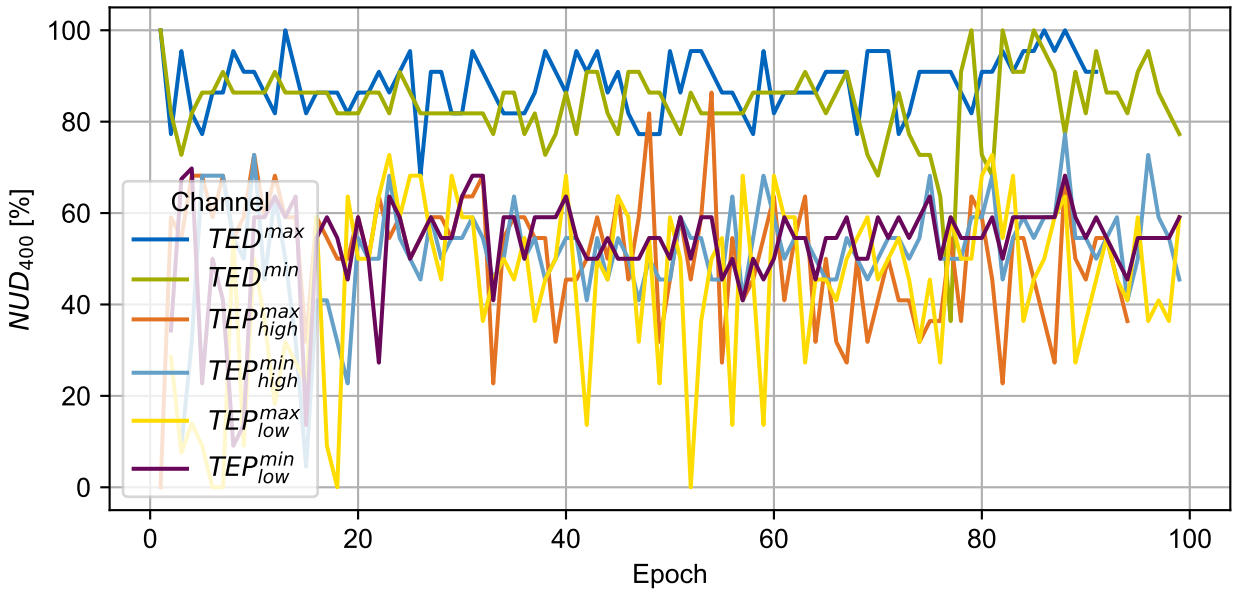


Figure 5.61 Number of false negatives larger than 400 µm (NUD_{400}) in specimen B23 per epoch for all six CNNs trained on individual channels.

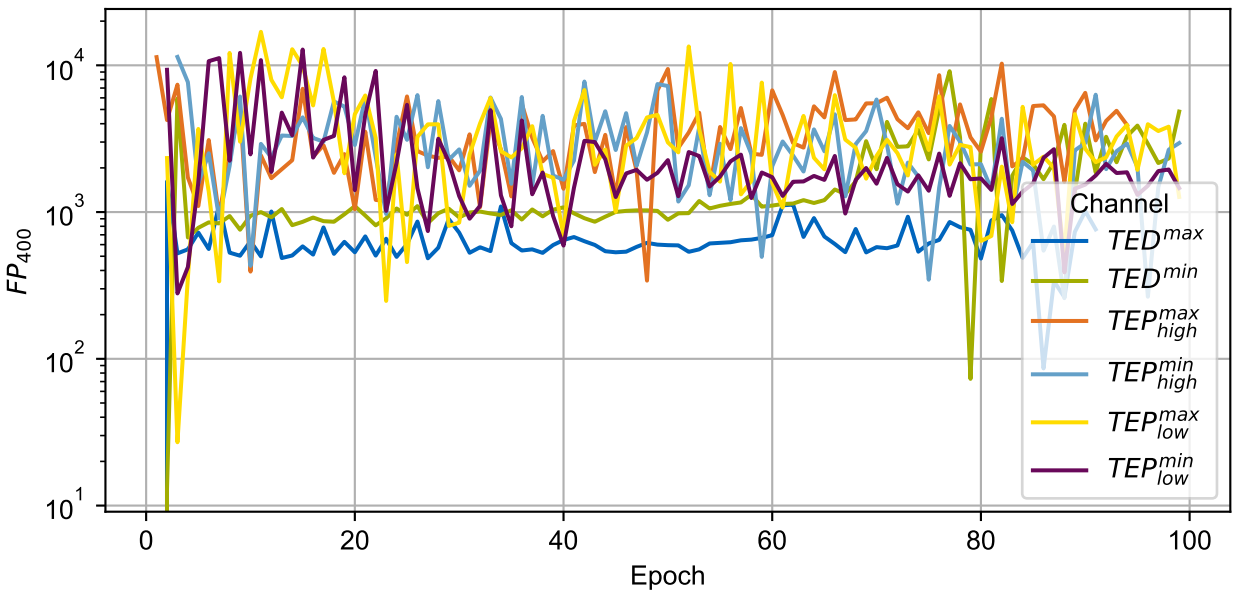


Figure 5.62 Number of false positives larger than 400 µm in specimen B23 per epoch for all six CNNs trained on individual channels (plotted logarithmically).

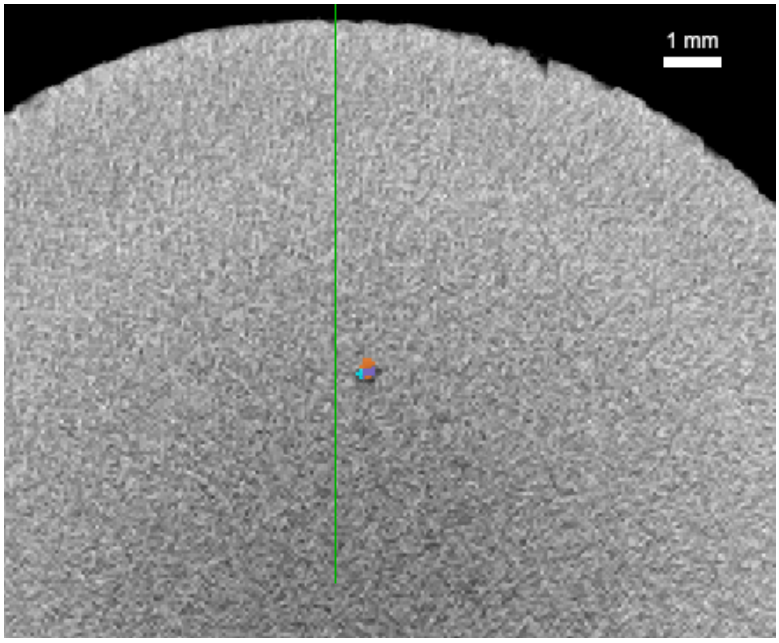


Figure 5.63 Exemplary inference result for specimen B23 of the single channel model (TEP_{high}^{min}, Epoch 88) in orange and the Ground Truth in blue. The overlap of both label maps is visualized in purple. The network seems able to produce meaningful results, but the overall performance lags behind the performance for samples A11 and B17.

signal-to-noise ratio caused by observing a wide bandwidth. This does not mean that data captured by the TED channel can be neglected entirely. It might contain valuable process information when combined with additional channels. To investigate this further, the results of the models trained on individual channels and the same model trained on all six channels are compared in the following. For this, the identical model architecture and training setup are used, except for adapting the number of input channels (one vs. six). For clarity, the quantitative comparison is based on specimen A11, as the single-channel networks are only trained on samples from Buildjob A. The other two qualified samples are used for qualitative comparison.

To allow for better visualization of the comparison, the networks trained on the single channels TED^{max}, TED^{min}, TEP_{low}^{max}, TEP_{low}^{min}, and TEP_{high}^{max} are neglected. Instead, the single-channel network trained on TEP_{high}^{min} is chosen to represent the performance of the single-channel models. It shows significantly better results than models trained on TED and represents the general behavior of the other models well. The metrics BUD, NUD_{400} and FP_{400} for A11 are plotted for both models in Figure 5.64, Figure 5.65 and Figure 5.66 respectively. During the training process, the BUD shows similar trends for both models. Only for Epoch 98, the BUD for the model trained on six channels drops shortly to around 568 µm while the BUD for the other epochs and TEP_{high}^{min} stays around 782 µm. When comparing the NUD_{400}, the model trained only on TEP_{high}^{min} demonstrates a better trend over almost all epochs. In contrast, the number of false positives larger than 400 µm shows an inverse trend. Here, the network trained on all six channels performs slightly better over the training process.

Summarizing the observations of the three metrics, the differences between the model trained on six channels and the model trained on only one channel do not show a clear picture. The only significant difference is the BUD for the respectively best epoch (568 µm for six channels vs. 782 µm for single channel). Upon detailed investigation, this BUD for the single channel model is identical to the BUD of the 1. OM U-Net (see Section 5.2.1 and Figure 5.27.) As described in these sections, such anomalies have to be regarded as ambiguous border cases. Considering the detection rate of the other relevant anomalies (NUD_{400}) and the number of relevant false positive predictions (FP_{400}), no clear improvement is observable from the single- to the six-channel network. Therefore, adding additional channels as input information does not show significant benefits for the investigated sample. This indicates that one channel (TEP_{high}^{min}) contains sufficient information for the CNN to predict defects.

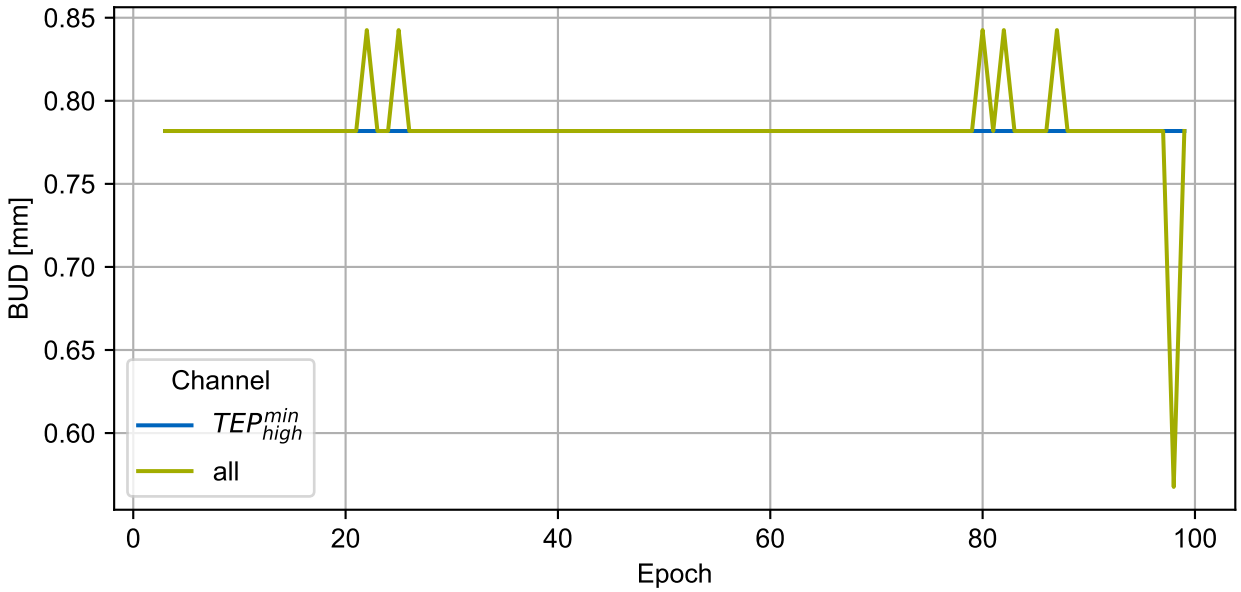


Figure 5.64 Comparison of the BUD in A11 for the model trained on only one channel (TEP^{min}_{high}) (blue) and the model trained on all six channels (green).

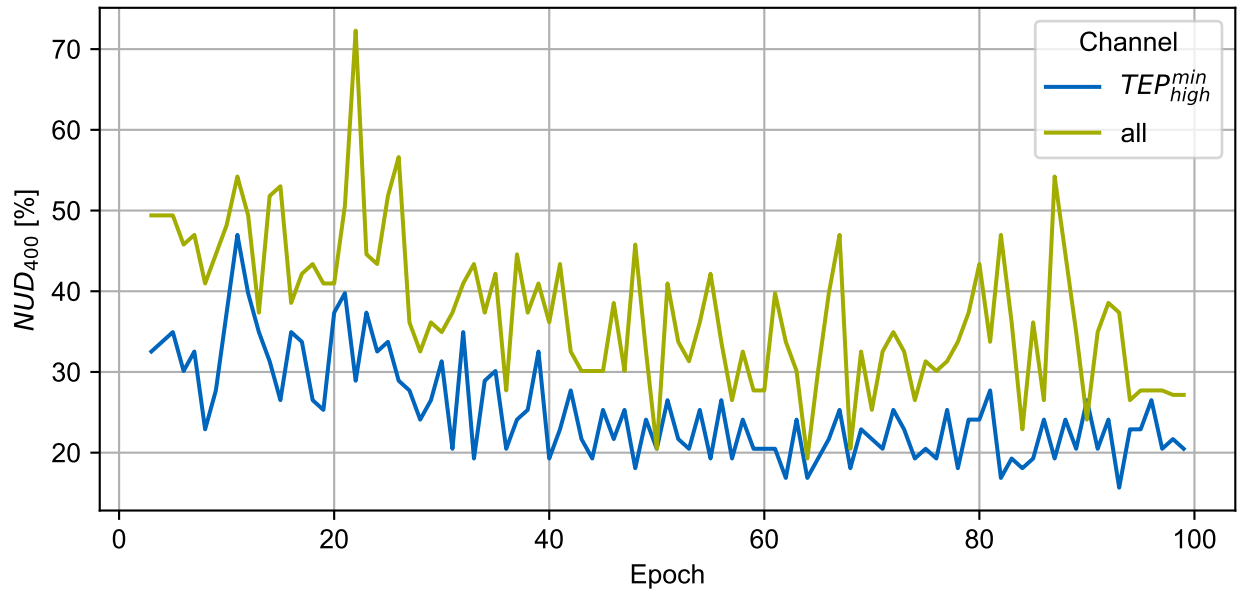


Figure 5.65 Comparison of the number of false negatives larger than 400 µm (NUD_{400}) in A11 for the model trained on only one channel (TEP^{min}_{high}) (blue) and the model trained on all six channels (green).

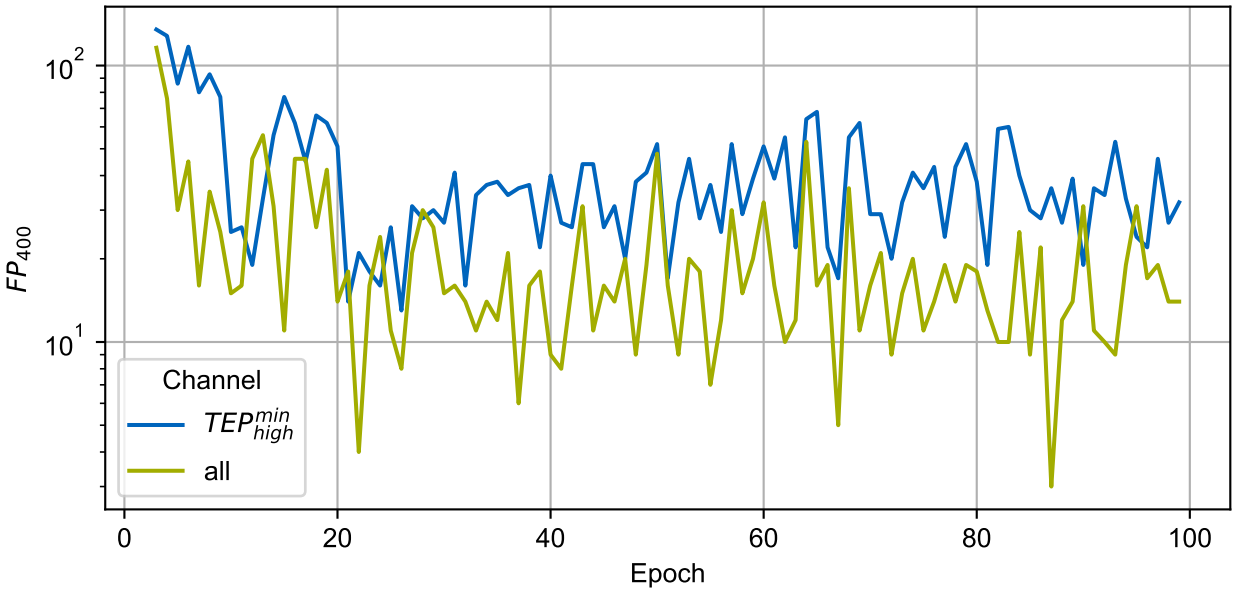


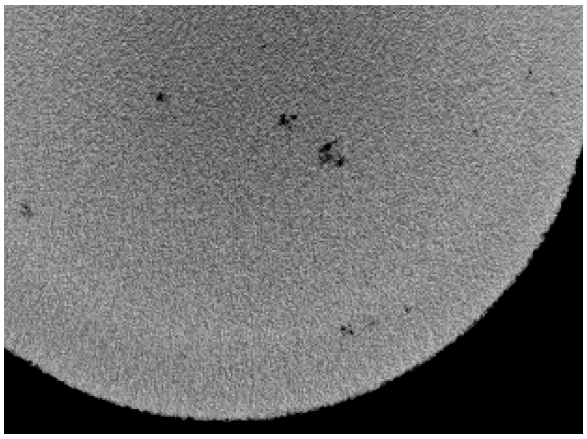
Figure 5.66 Comparison of the number of false positives larger than 400 µm in A11 for the model trained on only one channel (TEP^{min}_{high}) (blue) and the model trained on all six channels (green) plotted logarithmically.

Important to note here is that this observation is based on sample A11 and, hence, on one defect creation parameter and geometry. The models are trained only on specimens from Buildjob A. Therefore, when comparing the performance of the models on the test samples B17 and B23, a possible lack of transferability has to be considered. A visual, qualitative inspection (see Figure 5.67) of the inference results on B17 indicates a transferability of both models to other defect creation parameters. Furthermore, the visual differences between the single-channel and six-channel models are limited. Hence, suggesting that the information contained in the TEP^{min}_{high} channel is sufficient, and adding further channels does not significantly improve transferability to other defect creation types. The general transferability between defect creation modes will be investigated in detail in Section 5.2.4.

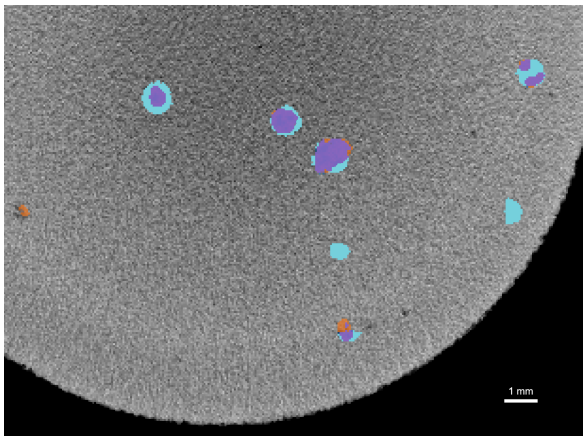
5.2.4 4. OM U-Net: Defect Modes

In the following, the results of the 4. OM U-Net are analyzed with a focus on transferability between different defect creation modes. As described in Section 4.6.4, this approach aims to analyze how the network performs on defects created by process deviations not included in the training data. If the network performs well, this indicates a good generalization to different defect provocation modes and can increase trust in the model results.

Training results The training loss curve in Figure 5.68 does not differ significantly from the previously described trainings. It indicates a quickly adapting network within the first 5000 steps. After that, the training loss fluctuates at around 0.25. The validation loss in Figure 5.69 shows a qualitatively similar behavior but at a lower level of around 0.09. The validation Dice in Figure 5.70 supports the assumption of a well-tuned training process. It increases sharply within the first epochs and then fluctuates at around 0.3. It shows a steeper increase in the Dice score than the 2. OM U-Net, which is trained with the identical training configuration but both buildjobs (Buildjob A and B). The validation data for both models is identical. This points towards a faster adaption of the network during training. For a quantitative interpretation, the pseudo-test results are analyzed as follows.



(a) The CT image showing the biggest defect in B17.



(b) The biggest defect overlaid with the inference of single- (orange)/ six-channel (blue) model. The overlap of both label maps is visualized in purple.

Figure 5.67 Visual inspection of the inference for B17 to qualitatively evaluate the transferability of models trained on only one channel (Epoch 96, orange) and six channels (Epoch 3, blue).

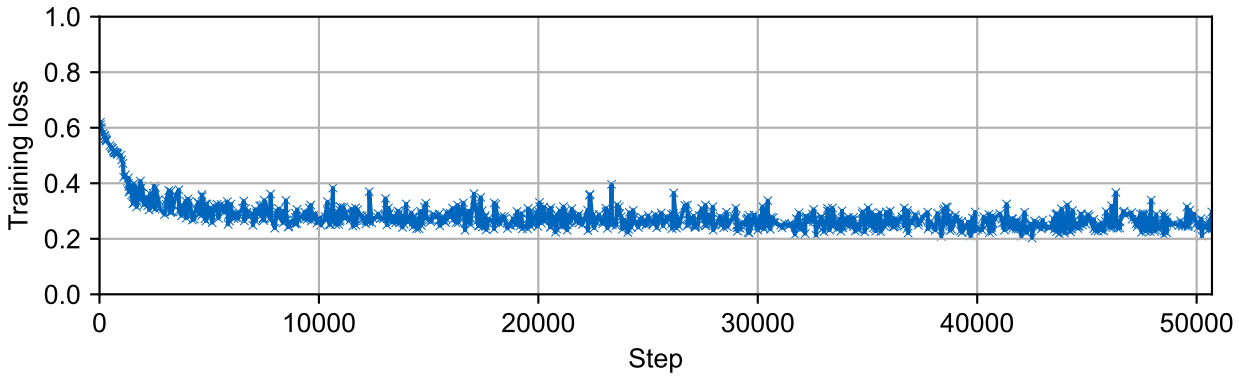


Figure 5.68 Training loss for the 4. OM CNN per training step.

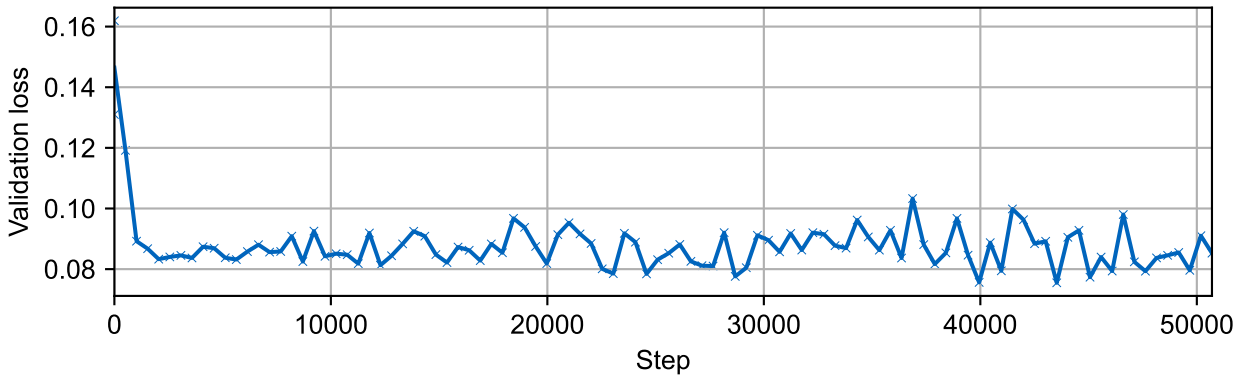


Figure 5.69 Validation loss for the 4. OM CNN per training step.

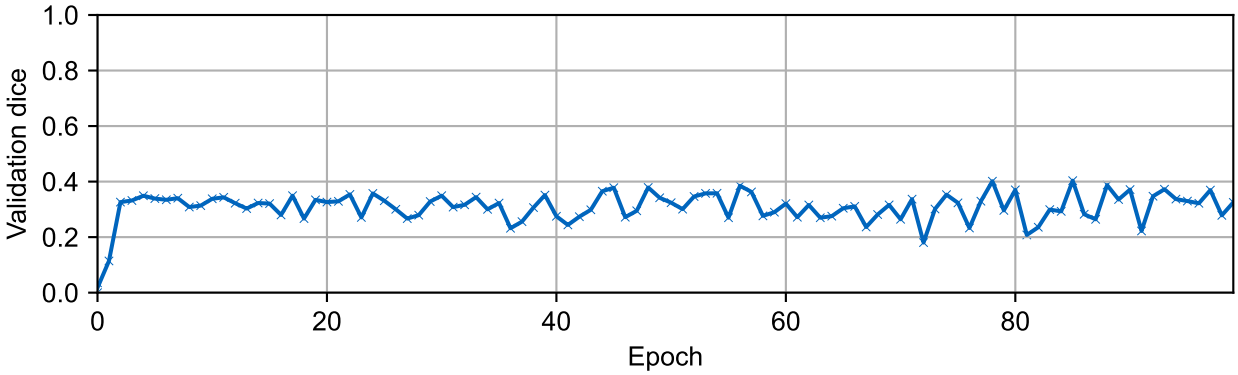


Figure 5.70 Validation Dice for the 4. OM CNN per training epoch.

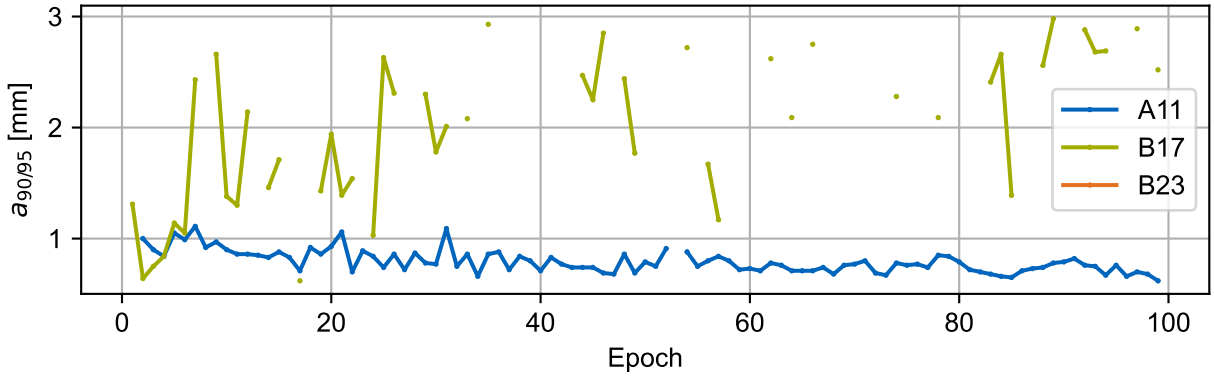


Figure 5.71 The POD per epoch for the 4. OM U-Net trained only on specimens from Buildjob A.

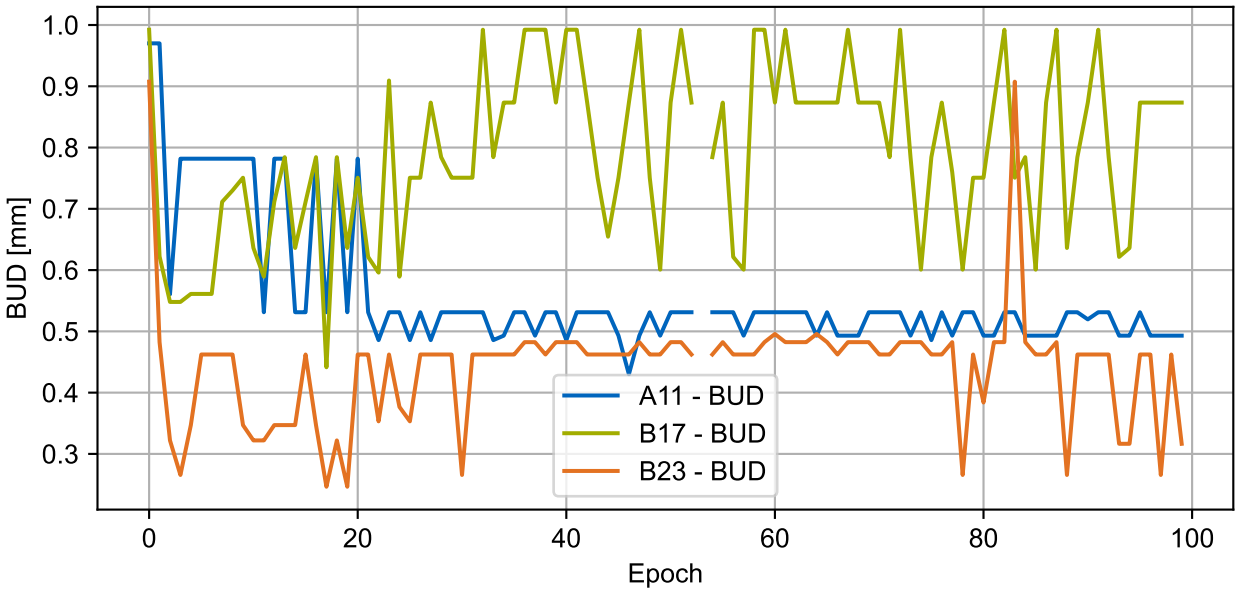


Figure 5.72 The biggest undetected defect per epoch for the 4. OM U-Net trained only on specimens from Buildjob A.

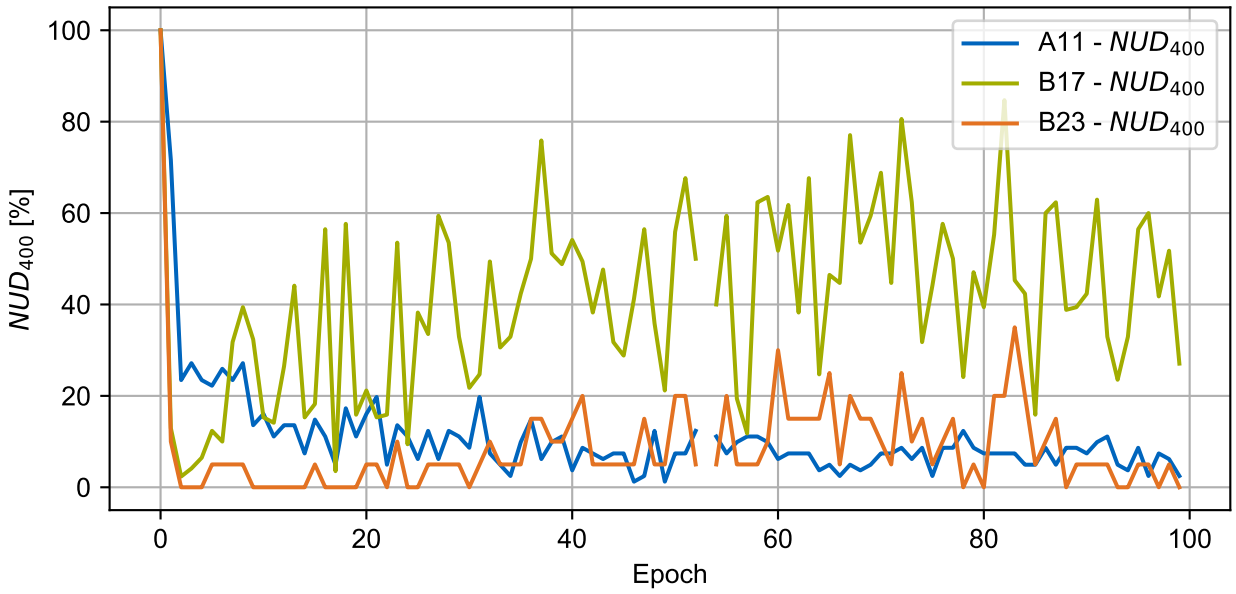


Figure 5.73 Percentage of false negatives larger than 400 µm (NUD_{400}) per epoch for the 4. OM U-Net trained only on specimens from Buildjob A.

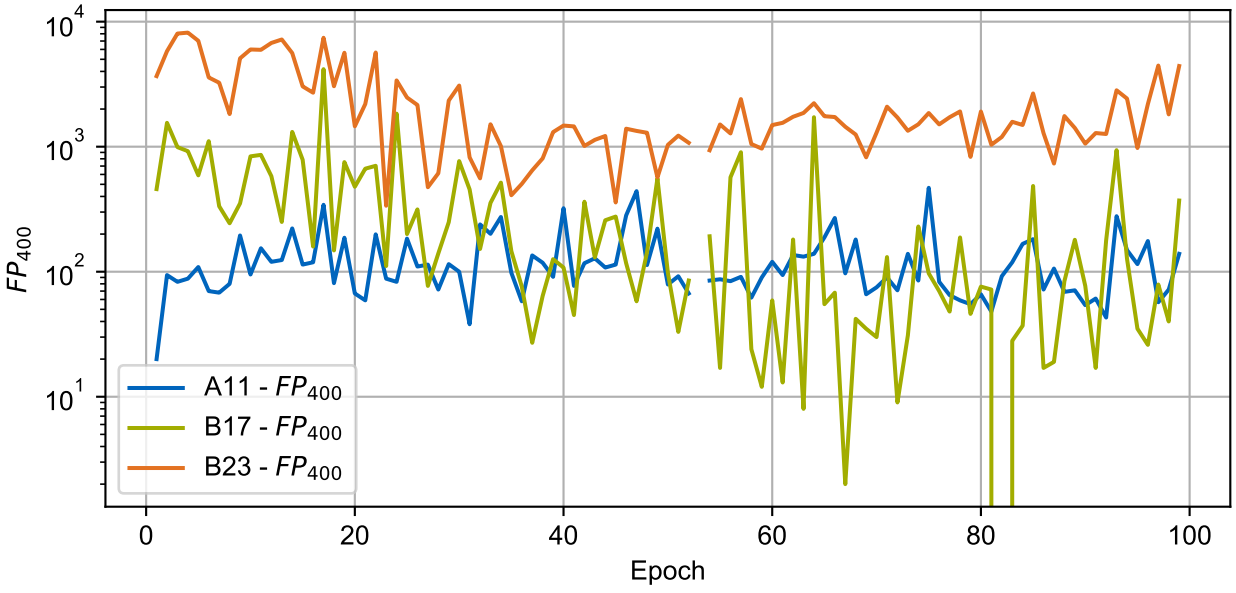
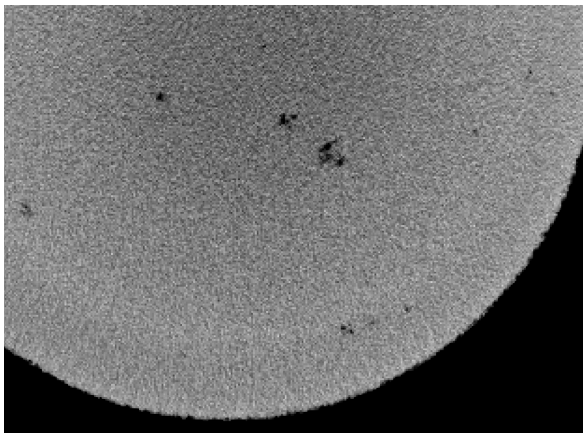


Figure 5.74 Number of false positives larger than 400 µm per epoch for the 4. OM U-Net trained only on specimens from Buildjob A (plotted logarithmically).

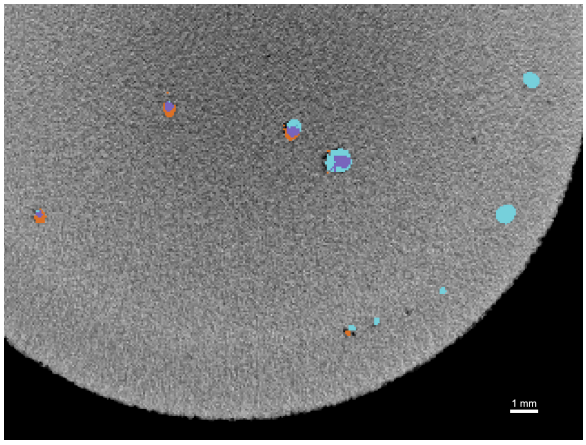
Pseudo-Test results The test results for the three specimens show a diverse picture. The POD (Figure 5.71) for A11 exhibits a similar trend as for the 2. OM U-Net with a fluctuation between 1.1 mm and 0.62 mm. The POD for B17, on the other hand, shows a significantly stronger fluctuation between 2.98 mm and 0.62 mm. For B23, no POD can be calculated for any epoch. The BUD (Figure 5.72) for A11 and B23 drop to around 500 µm with a relatively low fluctuation from Epoch 20 on. The BUD for B17 shows a stronger fluctuation over the entire training process. The best BUD for A11 is 430 µm at Epoch 46. The best BUD for B17 and B23 are reached at Epoch 17 with 441 µm and 246 µm respectively. The NUD_{400} (see Figure 5.73) shows a qualitative similar behavior to the BUD. The trade-off for the low BUD/ NUD_{400} at Epoch 17 is an extremely high false positive count (FP_{400}) of 343 (A11), 4169 (B17), and 7434 (B23) (see Figure 5.74). A visual inspection of the inference results confirms this. As a good trade-off between a low BUD/ NUD_{400} and a low FP_{400} for all three test specimens, Epoch 49 is identified. Based on this epoch and the results from Section 5.2.2, the influence of the training data and the generalization of the model is analyzed in the following.

Interpretation The main objective of the subsequent analysis is the testing of the two hypotheses stated in Section 4.6.4. Firstly, the model might detect defects based on process signatures produced by the actual defect creation process. Or secondly, the model detects defects based on the influence of defects on later layers. The model is trained on Buildjob A. All defect-prone specimens in Buildjob A are produced with the same parameter set. This parameter set is prone to create lack of fusion by a sub-optimal laser power. It is assumed that defects created by this parameter set exhibit a similar formation behavior. Hence, the defect formation signature and the signature produced by the defect in later layers should both be similar for all specimens in Buildjob A. Therefore, it is to be expected that the model performs well on A11 as it can detect defects in samples from Buildjob A for either hypothesis. This assumption is confirmed when comparing the test results with those of Section 5.2.2. The BUD (493 µm) is the same for both models (for the selected epochs). When comparing the NUD_{400} the 4. OM U-Net performs slightly better with a NUD_{400} of 1,2% (vs. 2,4% for the 2. OM U-Net). The FP_{400} is slightly worse, with 221 for the 4. OM U-net, compared to 190. Additionally, the POD also indicates a similar performance for both networks with a POD of 0.63 mm for the 2. OM U-Net and a POD of 0.69 mm for the 4. OM U-Net. Taking all four metrics into account, no significant differences can be observed. In combination with the visual inspection of the inference maps, it is concluded that the training on only Buildjob A has no significant effect on the prediction capabilities for A11.

Testing the model for Epoch 49 on the qualified specimen B17 results in a POD of 1.77 mm, a BUD of 601 µm and a NUD_{400} of 21%. In comparison, the selected Epoch 91 for the 2. OM U-Net has a POD of 0.75 mm, a BUD of 492 µm and a NUD_{400} of 3%. The FP_{400} is similar for both models 559 (4. OM U-Net) vs. 424 (2. OM U-Net). This indicates a significant decrease in performance for the model when trained only on Buildjob A. While the model's performance decreases compared to the model trained on Buildjob A and B, the model can still detect a significant number of defects in a relevant defect size range. The visual inspection of the label maps supports this observation. Figure 5.75 shows the slice (Slice 516) with the largest defect in B17. The CT image is shown on the left. The image on the right shows the CT image together with the label map for the 2. OM U-Net in blue and the label map for the 4. OM U-Net (Epoch 49) in orange. The overlay of both label maps is visualized in purple. The visual inspection shows good results for the largest defect and also for adjacent defects in the same slice for both models. Together with the quantitative metrics, the qualitative inspection indicates that the model is able to produce meaningful results even if not trained on Buildjob B. The deterioration of the quantitative evaluation might be attributed to a variety of factors. On the one hand, it cannot be ruled out that the inferior performance is partially caused by excluding a specific defect creation mode from the training data. On the other hand, the decreased number of training samples in itself might hinder the training process. Independently from the creation mode for defects, a reduction in training data might decrease the representativity of the training dataset and hence impede the training of the network. Secondly, the simple selection of the epochs based on a trade-off between false positives and false negatives results in differing performance metrics. Thirdly, even though the training configuration for the networks is kept constant, the



(a) The CT image showing the biggest defect in B17.



(b) The biggest defect in B17. With the label map of the 4. OM U-Net (Epoch 49) marked in orange, and the label map of the 2. OM U-Net (Epoch 91) displayed in blue for comparison. The overlap of both label maps is visualized in purple.

Figure 5.75 Visual comparison of the inference for B17 to qualitatively evaluate the transferability of the model to other defect creation modes.

training process differs slightly due to the different number of training samples, which might result in slight performance variations. In conclusion, even though the performance stays behind the reference model (2. U-Net), a basic transferability of the model is assumed for B17 based on the meaningful results produced for a majority of defects.

The inference for B23 results in a BUD of 462 µm, NUD_{400} of 5% and a FP_{400} of 568, compared to 908 µm, 59 % and around 750 for the 2. U-Net (Epoch 91). Therefore, the model performs better than the reference model on B23. As described above, the metrics for B23 have to be considered with extra care due to the high number of ambiguous cases (i.e., pores open to the surface) and the potential concurrence of defects with hatch lines. Hence, a direct quantitative comparison has to be supported by a qualitative interpretation of the inference (see Figure 5.76 and Figure 5.77). The visual inspection of the inference label map shows systematic false positive predictions at the outer border of the specimen for the 4. OM U-Net (Epoch 49). These outer false positives account for the majority of all false positive predictions. As they lie outside of the sample they could easily be removed by a post-processing of the labels (i.e. only keep labels inside of the sample). Meanwhile, internal false positives are reduced in comparison to the 2. OM U-Net. Furthermore, when compared to the 2. OM U-Net, the inspection of internal defects shows a similar qualitative detection performance for both models. Overall, a clear comparison of the two models for B23 is challenging. Nevertheless, the quantitative and qualitative evaluation of the results of the 4. OM U-Net indicate a basic transferability between defect creation modes with limitations.

Taking the findings from A11, B17, and B23 into consideration, the model performance indicates a basic transferability from Buildjob A to Buildjob B. While the BUD, NUD_{400} and FP_{400} deteriorate for B17 in comparison to the 2. OM U-Net, they improve for B23. The visual inspection of all inference label maps shows meaningful predictions with a good qualitative fit between Ground Truth and inference for actual defects.

In conclusion, the results of the 4. OM U-Net support the hypothesis that the model detects defects based on their influence on later layers. While it cannot be ruled out that the individual defect creation signature is also relevant to the network, the basic transferability indicates that the model is able to detect defects independently from their creation parameter. This is also supported by further studies concerning the explainability of the CNN presented in Milcke [2022] and Chapter 6. To investigate the behaviour of the meltpool further, it is planned to produce specimens with artificial defects at pre-defined locations such that their influence on later layers can be studied more precisely.

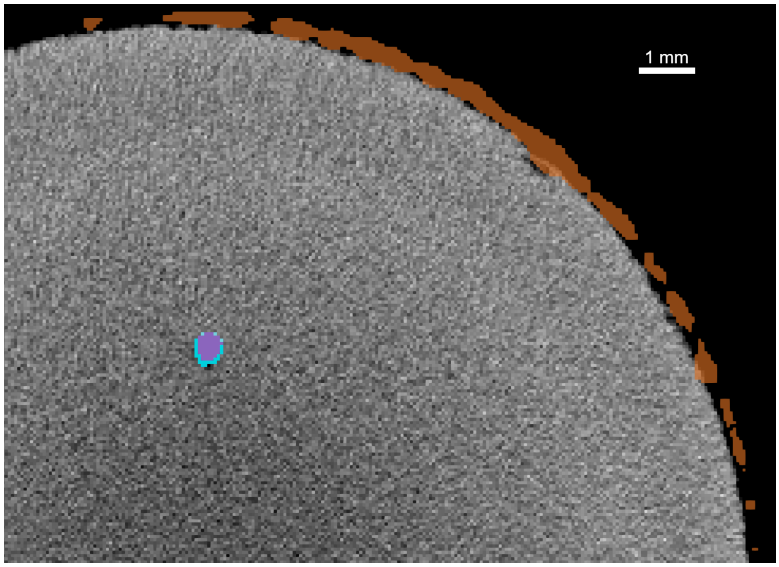


Figure 5.76 Exemplary defect in the specimen B23. With the label map for the 4. U-Net (Epoch 49) marked in orange and the label map of the 2. U-Net (Epoch 91) displayed in blue for comparison. The overlap of both label maps is visualized in purple. The clear systematic false positive prediction is visible around the samples. The label map for actual defects inside the specimen shows a good correlation with the GT and the prediction of the 2. OM U-Net.

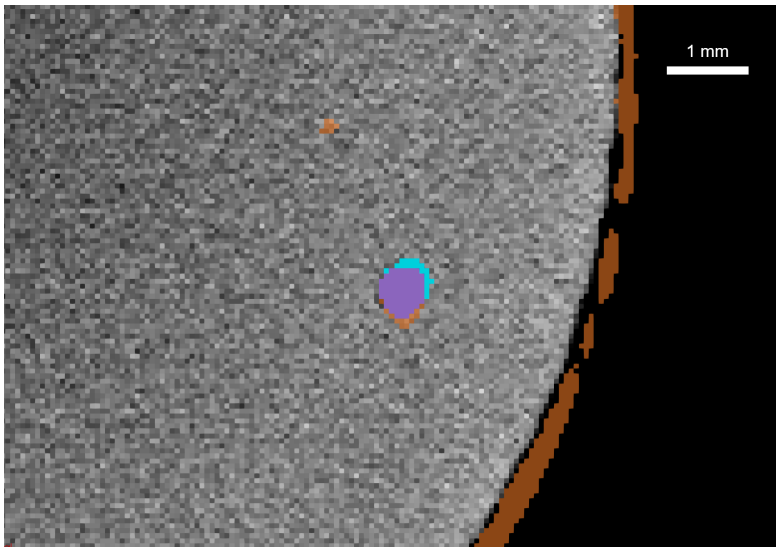


Figure 5.77 Second exemplary defect in the specimen B23. With the label map for the 4. U-Net (Epoch 49) marked in orange and the label map of the 2. U-Net (Epoch 91) displayed in blue for comparison. The overlap of both label maps is visualized in purple. The clear systematic false positive prediction is visible around the samples. The label map for actual defects inside the specimen shows a good correlation with the GT and the prediction of the 2. OM U-Net.

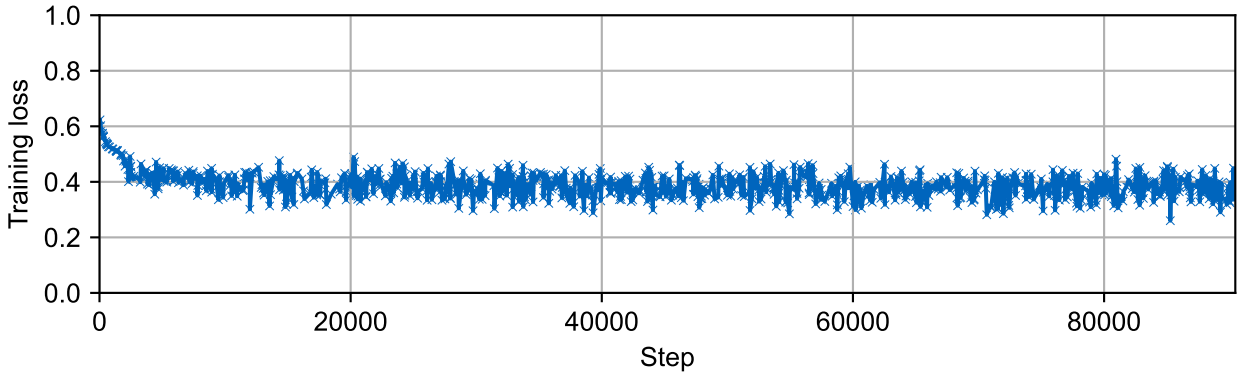


Figure 5.78 Training loss for the 5. OM CNN per training step.

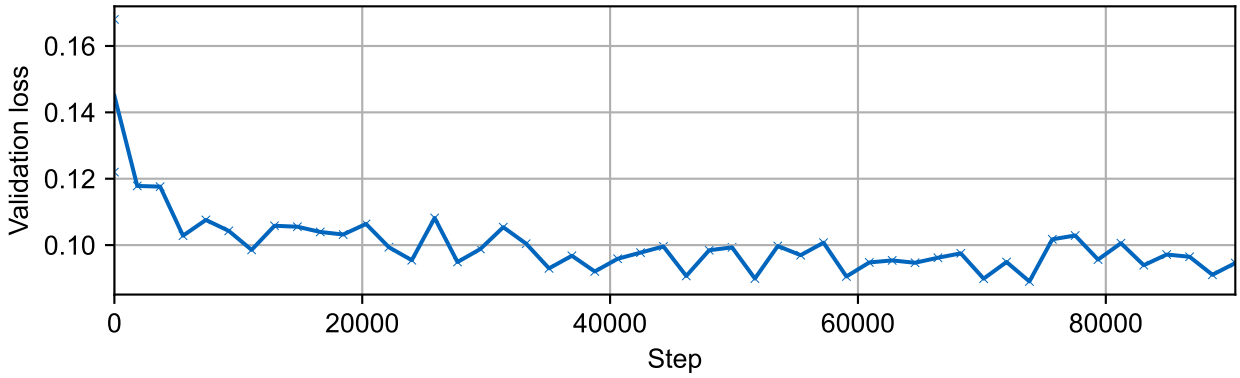


Figure 5.79 Validation loss for the 5. OM CNN per training step.

5.2.5 5. OM U-Net: Defect Size

The objective of the fifth training is the investigation of the influence of the training defect size. The motivation and training setup are described in Section 4.6.5. In the following, the network performance is analyzed with a particular focus on the detection of defects with sizes outside of the training size range. For training, only defects with a maximum bounding box length smaller than 500 µm are used, while the test data also contains defects significantly larger than 500 µm.

Training results The training and validation loss in Figure 5.78 and Figure 5.79 show similar qualitative behavior to the previous trainings in Section 5.2.1 and Section 5.2.2. A quick drop within the first 3000 steps suggests a well-tuned training and quick adaptation to the training data. The validation Dice (Figure 5.80) shows a gradual improvement over the training process before fluctuating around 0.2. In combination, the training results do not show significant impediments to the network training due to the artificially limited training basis. The actual performance of the network with respect to different defect sizes will be investigated in the following.

Pseudo-Test results The POD, BUD, NUD_{400} and FP_{400} for each test specimen per epoch are plotted in Figure 5.81, Figure 5.82, Figure 5.83 and Figure 5.84 respectively. The BUD trend for all three models already shows promising indications for the transferability of the model for different defect ranges. Even though the model is trained solely on defects smaller than 500 µm, it also detects defects larger than 500 µm. This is also supported by the POD, which can be calculated for most epochs for A11 and B17, indicating that the prerequisite for the calculation is fulfilled. The overall performance stays behind the

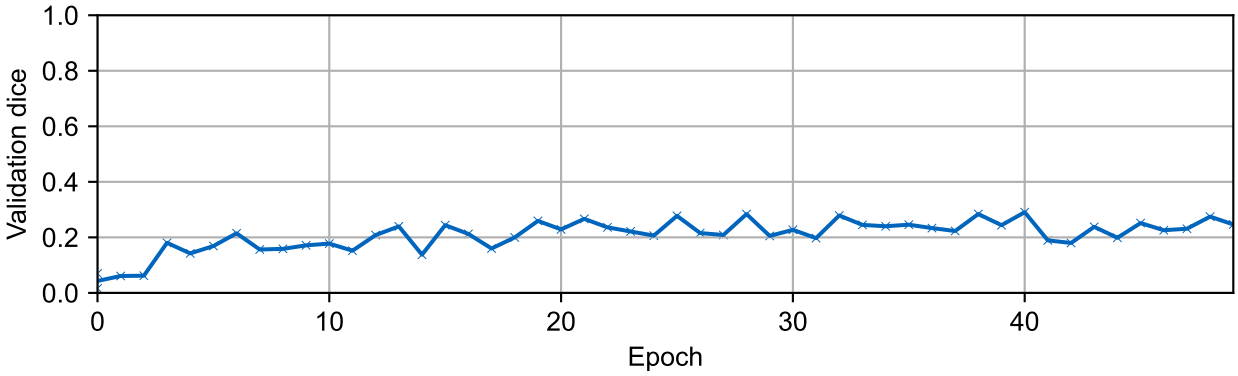


Figure 5.80 Validation Dice for the 5. OM CNN per training epoch.

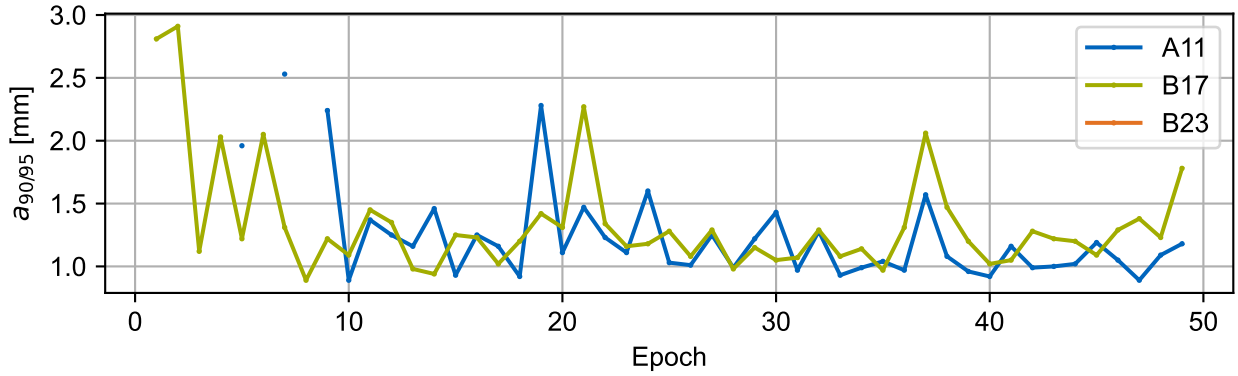


Figure 5.81 The POD for each specimen per epoch for the 5. OM CNN.

2. OM U-Net, but this might also be due to the reduced number of training cubes and is not necessarily solely attributed to the limited size range. For clarity, the transferability is evaluated in the following on the specimen A11. The best BUD for A11 is reached at multiple Epochs with 782 µm. In combination with a relatively low POD of 0.89 mm, a NUD_{400} of 34% and a low FP_{400} of 7 the Epoch 47 is chosen for further analysis.

Interpretation To analyze the model's generalization over a broad range of defect sizes, the performance of the model for Epoch 47 on defects larger than 500 µm is evaluated. The specimen A11 contains a total of 35 defects larger than 500 µm. Of those 35 defects, 30 are detected by the model. The three largest defects overall and the largest not detected defect (782 µm) are shown in Figure 5.85. The BUD is the same as for the 1. OM U-Net and represents an ambiguous label. The overall three largest defects have a maximum bounding box length of 970 µm, 843 µm, and 836 µm. Therefore, they lie well outside the training defect size range. Besides being able to detect those defects, the model also produces a meaningful label for them. The inference results correspond qualitatively to the orientation and rough size of the defects. In combination, the high detection rate for defects larger than 500 µm and the qualitative correspondence of inference and defect label map indicate a good transferability of the model to larger defects. Hence, it is assumed that the available data provides a sound basis for training and that the model is able to generalize the training data to larger defects. This is an important finding for the applicability of the POD, as the POD calculation requires a correlation between the defect size and the probability of the defect being detected. The 5. OM U-Net suggests that the CNN is not limited significantly in its performance when trained on a subset of smaller defects. As it can be assumed that the sensor hardware is more sensitive to larger defects, these results suggest that the overall system then fulfills the required correlation as well. To

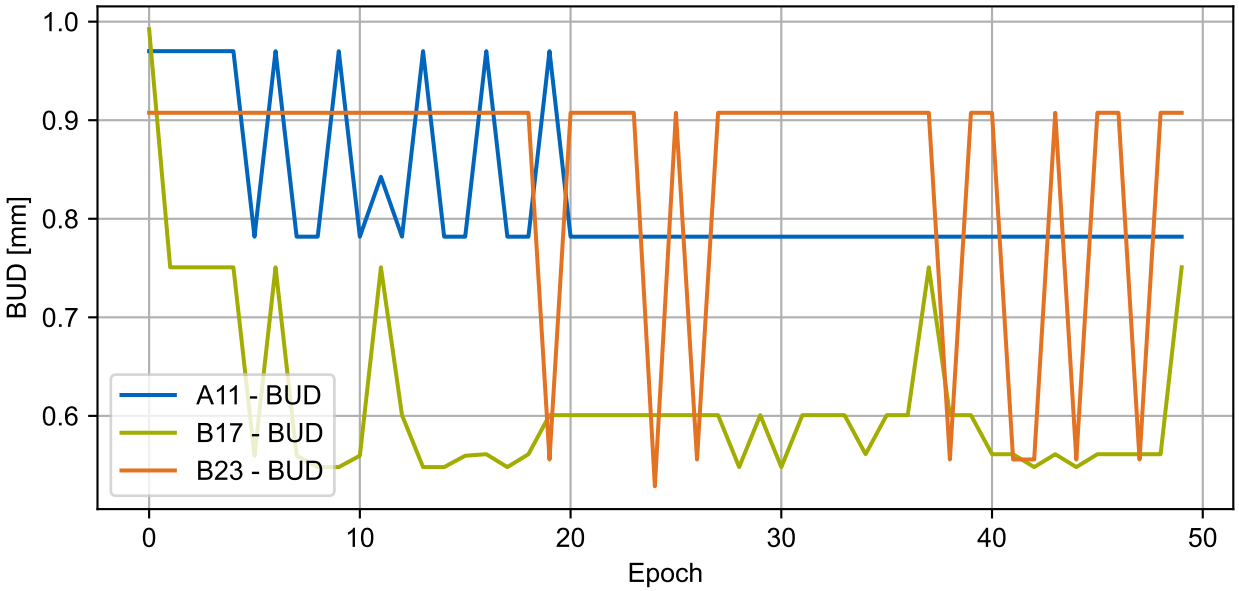


Figure 5.82 The biggest undetected defect for each specimen per epoch for the 5. OM CNN.

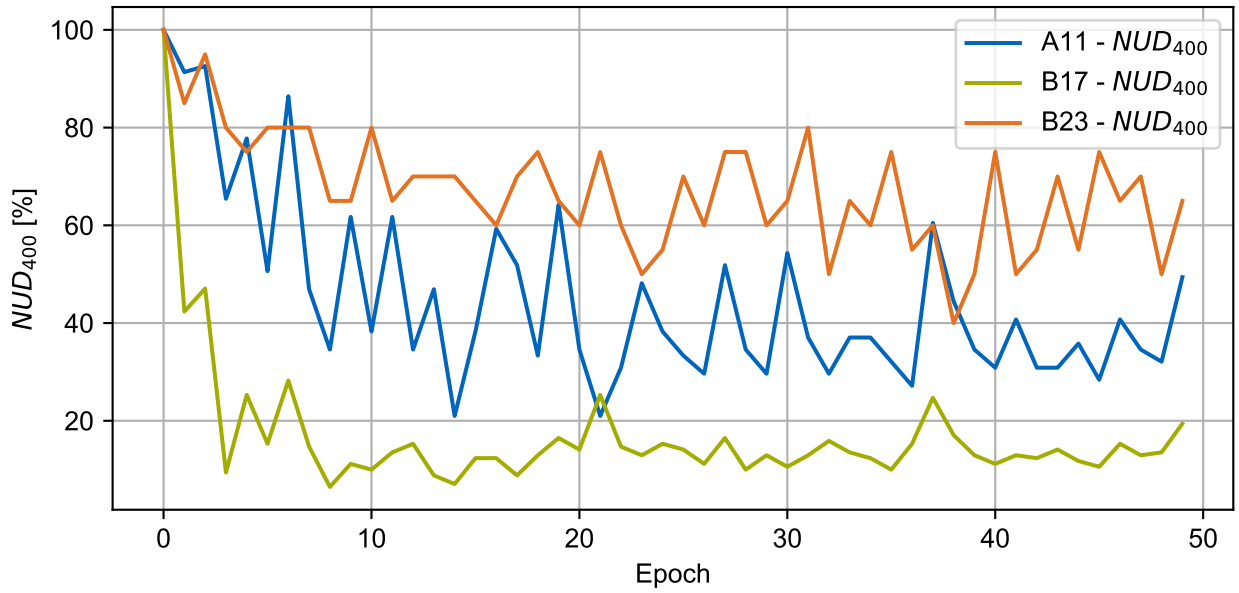


Figure 5.83 The percentage of undetected defects larger than 400 µm for each specimen per epoch for the 5. OM CNN.

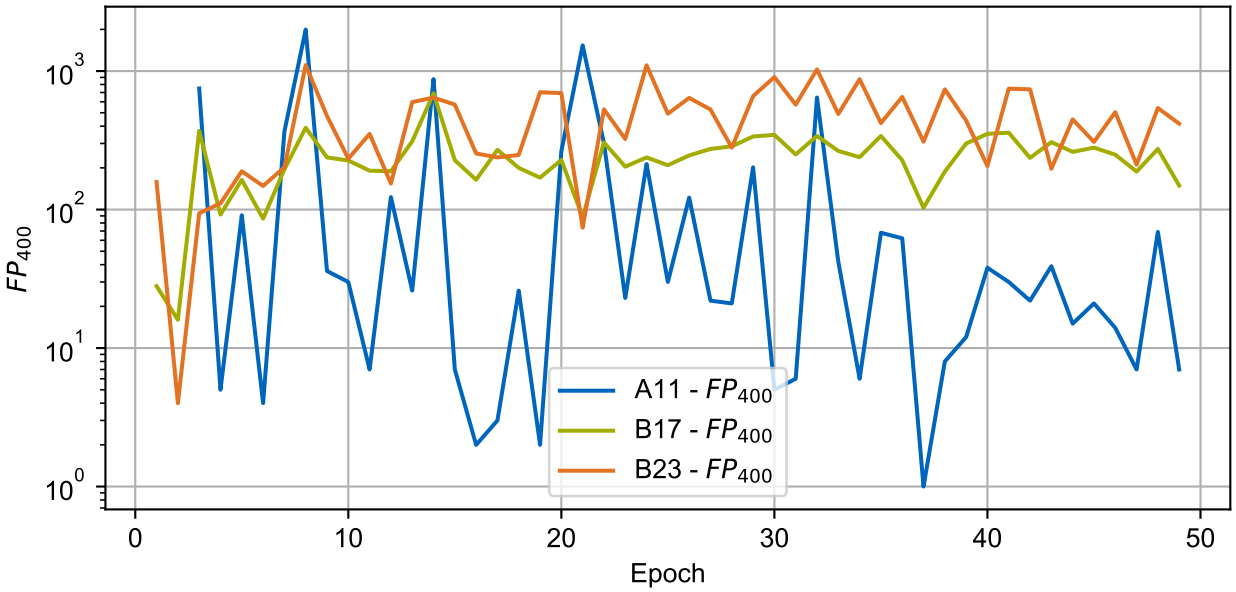
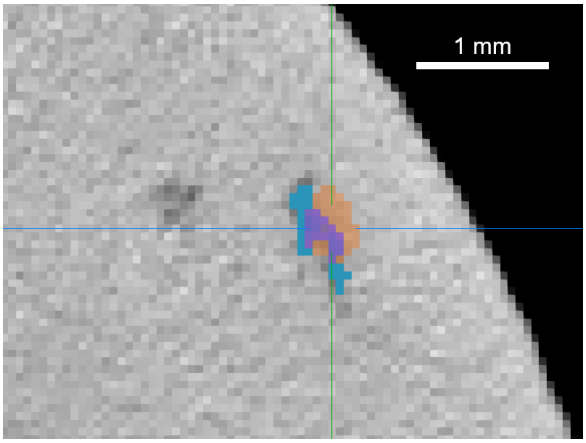
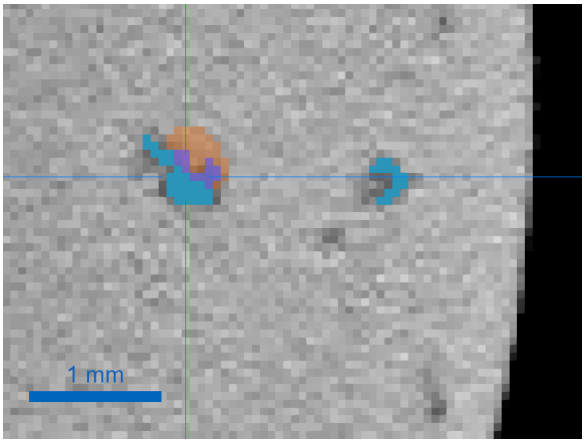


Figure 5.84 The number of false positives larger than 400 µm for each specimen per epoch for the 5. OM CNN.

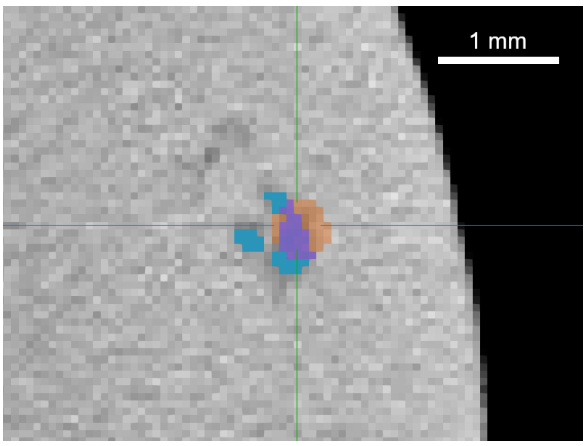
further analyze the generalization of the model, its performance on artificial data containing an even larger diversity of defects should be evaluated in future studies. Additionally, further specimens with even larger induced defects might support the current findings. This exceeds the scope of this thesis and is therefore proposed for further research.



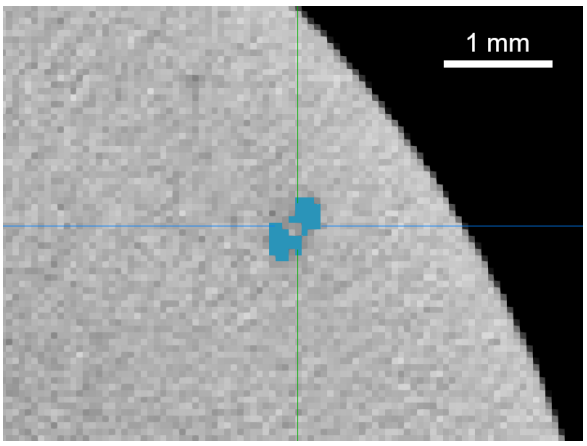
(a) The biggest defect overlaid with Ground Truth (blue) and the inference of the 5. OM U-Net at Epoch 47 (orange). The overlap of both label maps is visualized in purple.



(b) The second biggest defect overlaid with Ground Truth (blue) and the inference of the 5. OM U-Net at Epoch 47 (orange). The overlap of both label maps is visualized in purple.



(c) The third biggest defect overlaid with Ground Truth (blue) and the inference of the 5. OM U-Net at Epoch 47 (orange). The overlap of both label maps is visualized in purple.



(d) The biggest undetected defect overlaid with Ground Truth (blue).

Figure 5.85 The visual inspection of the largest defects in specimen A11 indicates a good transferability of the model to larger defect sizes than those included in the training dataset.

5.2.6 OM CNN Summary

The five presented experiments provide insights into different aspects of the proposed CNN and overall system. The first two CNNs (Section 5.2.1 and Section 5.2.2) show the general feasibility of the system to detect defects based on the online monitoring data. With a POD of around $0.63\,\mathrm{mm}$, the system shows great potential to reliably detect defects of relevant size. Additionally, the FP_{400} and qualitative visual inspection show the ability of the model to produce meaningful predictions. Here, the ambiguity of the qualified label map has to be taken into account (see Section 4.3.4).

The 3. OM U-Net (Section 5.2.3) investigates the importance of the individual sensors and channels. The results indicate that individual narrow bandwidth sensors (TEP_{low} and TEP_{high}) contain sufficient information to detect defects. In contrast, the models trained on the broad bandwidth sensor data (TED) show a significantly worse detection performance. As this photodiode captures the information of a broad spectrum, including the bandwidths measured by the narrow bandwidth sensors, it is assumed that the information about a possible defect is contained in the data but obscured by the additional noise captured in the broad spectrum. Therefore, the signal-to-noise ratio is regarded as significantly worse than for the narrow bandwidth signals. The comparison of the TEP_{high}^{min} model with the six channel model indicates no significant differences. This suggests that a single-channel hardware setup and CNN are sufficient, which should simplify the transfer to other monitoring systems.

The 4. OM U-Net (Section 5.2.4) focuses on the different defect provocation modes (keyhole vs. lack of fusion parameters). It analyzes the transferability of a model trained on Buildjob A to analyze defects in Buildjob B. Even though the network performance varies with regards to the 2. OM U-Net trained on both buildjobs, the visual inspection indicates a basic transferability between the different defect provocation modes. This suggests that the model can detect defects independently of their provocation mode. Further explainability approaches to support this hypothesis are presented in Milcke [2022] and discussed in Chapter 6.

The 5. OM U-Net (Section 5.2.5) analyzes the performance of the network for defect sizes not included in the training dataset. The experiment shows that the model is able to detect defects significantly larger than those contained in the training dataset. This suggests that the model generalizes well for larger defects. This is an essential prerequisite for using the POD, as the POD calculation requires a correlation between the defect size and the probability of the defect being detected.

The presented experiments with the main objectives, hyperparameters, and results are summarized in Table 5.5.

Table 5.5 Summary of the test results for the five presented OM CNNs.

	Main Objective	Hyperparameters	Result
1. OM CNN	Baseline & Feasibility	baseline parameters to establish general feasibility	basic feasibility of the proposed system is shown with a POD of around $0.78\,\mathrm{mm}$ (A11)
2. OM CNN	Feasibility & Performance Optimization	enhanced data augmentation	performance of the CNN is improved to $0.63\,\mathrm{mm}$ (A11) by enhanced data augmentation
3. OM CNN	Individual Channels	individual sensor channels	prediction of defects is feasible for networks trained on single narrow bandwidth sensor data
4. OM CNN	Defect Mode	training data limited to Buildjob A	basic transferability and explainability study indicates that network focuses on influence of defect on later layers
5. OM CNN	Defect Size	training data limited to defects smaller $0.5\,\mathrm{mm}$	detection of larger defects indicates basic transferability to larger defects

6 Discussion

In the following, the results obtained in Chapter 5 and additional studies are summarized and discussed with regard to the overall aim of this thesis. As stated in Section 1.2, this thesis' objective is the research of a novel defect detection method based on online monitoring data and the use of CNNs. For this, the processing pipeline illustrated in Figure 4.1 is designed. At the core of this pipeline is the training of the CNNs (Chapter 4 and Chapter 5) for defect detection in CT data and online monitoring data. Besides the proof of feasibility, additional studies are highlighted to improve the understanding and trust in the developed CNNs.

6.1 Computed Tomography

As discussed in Section 5.1, the CT CNN is trained in an iterative manner. The first ground truth label map is created semi-automatically with considerable manual interaction (see Section 4.3). Based on this label map, a first CNN is trained and subsequently used for inference to create additional label maps. By manually optimizing these label maps, a refined input for the subsequent CNN training is possible. Qualified CT inspectors refine a selection of the label maps. This approach allows for the creation of qualified GT data with reasonable effort.

While the first presented model has a low POD (315 µm) and low BUD (465 µm), it produces a high number of false positives (1775). Furthermore, the visual inspection of the label map reveals two possible systematics for false positives, which are attributed to CT artifacts. The second CT CNN reduces those systematics and the FP count by fine-tuning the training data and training process. By introducing part of the qualified label maps to the training data, the third CT CNN further decreases the false positive predictions while maintaining a low POD (330 µm) and BUD (378 µm). The visual inspection of the results confirms the mitigation of the false positive systematics observed in the first label map. The third CT CNN provides a sufficient performance to reliably detect relevant defects in CT scans. Therefore, it is used to analyze the CT scans automatically for the subsequent OM CNNs training. This represents an important step in the pipeline introduced in Figure 4.1. The automatic and reliable labeling of large amounts of CT data is an essential prerequisite to developing the OM CNN, as it provides the necessary GT data.

Besides this use-case-specific application of the CT CNN as a GT generator, the investigation of the CT CNNs provides additional findings about the use of CNNs for the analysis of CT scans in non-destructive testing. Firstly, adopting the U-Net architecture from medicine to non-destructive testing produces promising results. The developed CNNs are able to reliably detect defects in high-resolution CT scans of simple titanium parts. The effects of CT artifacts caused by metal can be mitigated by improving the label map quality. Secondly, the importance of the training data on the network performance is highlighted. The network architecture and training configuration are only altered slightly for the three introduced CNNs. The performance improvements are mostly attributed to an improved training dataset. Thirdly, the creation of high-quality GT data is identified as an important task in itself. The iterative improvement of CNN and GT shows good results. The manual verification and enhancement of selected label maps create additional reliability to the labeling process, resulting in the qualified GT label maps with the limitations described in Section 4.3.4. The qualified label maps are excluded from the training data for the OM CNN and are solely used to evaluate the OM CNNs.

6.2 Online Monitoring

The general approach to the evaluation of the CNN results is split into three steps. Firstly, the well-established training and evaluation metrics in machine learning are used to analyze the training behavior of the network. Secondly, the newly introduced, use-case-specific metrics BUD, NUD_{400}, FP_{400}, and the POD are calculated per epoch to select well-performing networks. These metrics allow for a comparison of the networks with each other for the defined NDT use case. Thirdly, a visual inspection of the selected label maps provides additional information about potential systematic errors or the influence of the ambiguity introduced by the qualified label maps. In combination, the quantitative and qualitative evaluation of the predicted label maps provide an in-depth analysis of the network performance.

6.2.1 Feasibility

Literature and pre-studies show the need for using CNNs as these studies did not find suitable conventional image processing algorithms to automatically detect defects in OM data robustly [Schawaller, 2019; Kaehn, 2020]. The feasibility of the proposed approach of using CNNs is demonstrated by the first two U-Nets presented in Section 5.2. The **1. OM U-Net** (Section 5.2.1) lays the baseline for this study. It uses all six input channels from the installed photodiodes. This allows the network to select from all available monitoring data. The network is trained for a small number of epochs to establish a first baseline and investigate the general feasibility. The training and validation metrics suggest a well-adapting network. Based on the test metrics, Epochs 8 and 20 are selected as exemplary. For both epochs, the network detects all anomalies larger than 781 µm (BUD). The NUD_{400} and FP_{400} vary strongly between the specimens. Therefore, visual inspection is used to support the quantitative results of these epochs. It shows that the form and location of the predictions mostly correspond well with the defect characteristics as evaluated by the CT. The high variation is mostly attributed to the ambiguity in the qualified label map.

Overall, the 1. OM U-Net shows great potential for the automatic evaluation of AM parts. The CNN can detect defects in titanium specimens based on the data gathered during the process. More specifically, the algorithm is able to segment defective regions in the specimen on a voxel level using the radiation of the meltpool in different wavelengths. This demonstrates the general feasibility of the presented approach.

The **2. OM U-Net** (Section 5.2.2) aims to improve the network performance to allow for a more accurate detection of defects. While the network architecture is kept constant, substantial data augmentation is applied to create a larger and more diverse training dataset. This leads to a slower adaptation during training, as indicated by the slower improvement in the validation Dice score. At the same time, the enhanced training data results in an improved performance of the network for later epochs. For the Epoch 83, it detects all defects larger than 531 µm. In particular, the BUD for A11 drops strongly from 781 µm for the 1. OM U-Net to 493 µm.

The analysis of the NUD_{400} and BUD on the one hand and the FP_{400} on the other hand reveals an inverse correlation of the metrics. Epochs with a low NUD_{400} and BUD have a higher FP_{400} and vice versa. This trade-off between a high detection rate and a low false positive count is common in many areas of NDT and ML. The choice of which metric to focus on is use-case-dependent.

Two exemplary contrary scenarios are plausible for the OM CNN. Firstly, replacing the CT inspection with online monitoring. In this case, the quality inspection would rely entirely on the developed algorithm. Therefore, the focus should be placed on the POD, NUD_{400}, and BUD as they ensure that relevant defects are detected reliably. Secondly, as a first step for industrialization, the system could be employed to stop a buildjob during printing if it detects a defect. The CT would be used for the final quality inspection in this scenario. This would allow for a quick monitoring system implementation into production as the qualification efforts could be reduced drastically. In this case, the benefit of the system lies in saving resources and time, as a buildjob could be stopped as soon as a defect is detected. Here, the focus should be placed on the FP_{400} as a high number of false positive predictions would lead to a high number of unnecessary printing stops. Possible ways to consider the targeted use case include the adaptation of the probability threshold, tailoring the training data, and adapting hyperparameters during training (e.g., weighted loss function).

Overall, the 1. and 2. OM U-Net show the feasibility of the proposed approach to train a CNN to automatically analyze the gathered OM data. While the 1. OM U-Net lays a baseline, the 2. OM U-Net highlights the importance of data augmentation for the robust training of the U-Net for online monitoring.

6.2.2 Explainabilty

To improve the understanding and trust in the developed CNN, additional studies are conducted to focus on the explainability of the detection results. Firstly, the **3. OM U-Net** (Section 5.2.3) focuses on the influence of the individual sensor channels to improve the understanding of the sensor system and the information captured by it. The analysis indicates that the TED photodiode with a broad bandwidth filter has a significantly worse signal-to-noise ratio in comparison to the two narrow bandwidth photodiodes. Models trained solely on this photodiode show a significantly inferior performance to models trained on TEP_{low} or TEP_{high}. In contrast, the U-Nets trained on the photodiodes with narrow bandwidth filters (TEP_{low} and TEP_{high}) show a quick training process and good pseudo-test performance. The differences between TEP_{low} and TEP_{high} and the maximum or minimum channel are regarded as marginal.

Secondly, these findings are supported by a dedicated **explainability study** performed in the scope of this project [Milcke, 2022]. The study investigated explainability from three different perspectives: data explainability, model explainability, and prediction explainability.

The investigation of the OM and GT data (data explainability) supports the assumption that narrow bandwidth channels have a higher importance for the detection of pores. The intensity distribution of these channels shows a higher fluctuation for defect regions, whereas defect-free regions follow a normal distribution in intensity. Furthermore, the influence of the individual channels of the six-channel U-Net is investigated by altering the input channels and observing their impact on the detection performance (prediction explainability). It is concluded that one of the narrow bandwidth channels can contain sufficient information for detecting pores. This implies that the measurement of the radiation of the meltpool is sufficient for the investigated approach, and an exact measurement of other physical properties (e.g., the temperature) is not mandatory for the detection of a pore.

In summary, it can be assumed that the bandpass filtering of the meltpool radiation significantly influences the data quality. The limitation to a narrow bandwidth seems beneficial and sufficient for pore detection. The influence of the selected bandwidth on the evaluation cannot be determined as the filter setup is fixed to two wavelengths for this study. Additionally, the findings not only provide insight into the relevance of individual channels but also strengthen the trust in the developed model. It suggests that the model can distinguish well between relevant and irrelevant features in the input data. In particular, it is able to focus on relevant channels and discard others.

6.2.3 Transferability

In addition to the investigation of the different channels, the influence of the defect size and the provoked defect creation mechanism on the detection performance are analyzed based on the 4. and 5. OM U-Net and the explainability study. This allows for a better understanding of the transferability of the model to new data and its limitations.

The **4. OM U-Net** (Section 5.2.4) investigates the transferability of the model between different defect mechanisms. Showing that the model is able to detect defects provoked by parameter set B while only being trained on defects provoked by parameter set A suggests that the model does not focus on the pore formation process. Instead, the results indicate that it focuses on the signature created by the pore in later layers due to an altered heat flow in the specimen.

To investigate this behavior further, the explainability approach introduced above is extended to evaluate the spatial location of relevant information for pore detection [Milcke, 2022]. The visualization of the relevant voxels in the input data for the U-Net shows that the network focuses on voxels in close spatial proximity to the defect. More specifically, these voxels are located in a hull shape around the upper half of the defect and above. The same observation holds true for the visual inspection of selected input data regions in which features that might correlate to a defect occur above the defect. In particular, an increased

visual correlation of deviations in the image gradient of the input data with a spatially lower defect could be observed in some cases. Showing that the model focuses on areas in the upper part and above a pore is a first step in supporting the assumption that the model focuses on the interaction between the process and the existing pore. This should facilitate the training of CNNs as the training data only needs to include specified pore geometries and not specified pore creation scenarios, which would be necessary if the model detects defects based on their creation signature. For example, if the model focused on features created during the formation of a pore, all possible formation scenarios would have to be covered by the training data. By showing that the model can detect defects based on their influence on later layers, this might not be necessary. Restrictively, it should be noted that the systematic upward shift might also be caused by a systematic registration error due to imperfections in the alignment of the CT label map and the OM data introduced by the registration. This seems unlikely as this would not explain the basic transferability between different defect provocation types.

In conclusion, the current findings of the 4. OM U-Net and the explainability study by Milcke [2022] indicate a basic transferability of the model between defect creation modes and suggest that the model focuses on pore existence, not pore formation. Additional insights could be obtained by supplementary specimen designs and by linking the data analysis to physical phenomena, e.g., by a numerical simulation of the LPBF process, particularly the heat flux within the specimen for porosity.

The **5. OM U-Net** focuses on the ability of the model to generalize well over a range of defect sizes. The results in Section 5.2.5 show that the model is able to detect defects larger than those contained in the training dataset. This indicates a basic transferability of the model between different defect sizes. In addition, the explainability study concludes that a larger defect correlates to a stronger signal in the gathered data [Milcke, 2022]. The two findings suggest that larger defects are easier to detect if the model is adequately trained. This is of importance for two reasons. Firstly, the POD calculation presumes a correlation between the defect size and the probability of the system detecting it, i.e., a larger defect is more likely to be detected. Hence, the findings support the use of the POD as a performance metric. Secondly, training the network on all possible defect sizes is not feasible in practice. Instead, the model should be trained on a defined range and generalized to a larger defect size range. The diversity of the training data might be improved by data augmentation (i.e., spatial scaling) during the training.

6.3 Limitations

In the following, the main limitations and challenges of the introduced approach are discussed to provide a baseline for further research objectives in Chapter 7.

Firstly, the investigated defects are all induced artificially by altering the process parameters of the AM process, as the normal process does not create defects in sufficient numbers to train CNNs. Instead, the specimens are produced with sub-optimal process parameters. This leads to a significantly higher number and size of defects per specimen. Even though the defects are not induced by local variation of the process and should behave similarly to real defects, this could not be shown within this work due to a lack of "real" defects. Therefore, the transferability to real defects should be evaluated in future studies based on long-term data gathering in real-life applications.

Secondly, the CT CNN is not tested on more complex geometries and other materials. Hence, the provided performance metrics are only valid for the investigated specimen geometry and material. It can be assumed that the CT CNN performs inferiorly on more complex geometries with more CT artifacts or inferior scan quality. A transfer learning approach should allow for an efficient adaptation of the current CT CNN to additional specimens and materials.

Thirdly, the data could contain a bias towards the shape and number of pixels of a defect, as proposed by the CNN. The non-refined label map is created by the CNN and used as a basis for the inspection. Even though the inspector analyzes each layer individually, he/ she might tend to agree with the proposed label map and, therefore, incorporate prior behaviors of the CNN into the qualified label map. For the presented use case, the influence of this bias on the performance evaluation seems marginal, as the exact shape of the defect is of second-order importance to the pure detection of the defect.

Fourthly, the evaluation of the CT scan produces a number of difficult border cases. As described in Section 4.3.4 and Chapter 5, the current procedure for the creation of the qualified label map leads to some anomalies that do not have to be labeled according to regulations but can be labeled if spotted by the inspector. This behavior is difficult to represent in a rule-based logic and might impede the CNN training and evaluation. In particular, the quantitative evaluation of the CNNs on the described border cases might result in an inferior performance statement. Therefore, in the scope of this thesis, the quantitative interpretation of the results is supported by a manual, qualitative analysis of the respective anomalies. Further steps to improve the qualified label map and the quantification procedure might include labeling from scratch, labeling the same volume by multiple inspectors, and defining requirements more precisely with regard to relevant defects for this study.

Besides the limitations introduced by the specimen and its CT analysis, the registration process has to be considered a potential source of inaccuracy. The CT label maps have to be aligned with the OM data in order to create meaningful training, validation, and test data. The registration mechanism designed for this purpose aims to find the best rigid transform to align the CT volume with the OM volume. The two volumes do not fit perfectly, as they are created from different sources (CT scanner vs. photodiodes). This introduces an inaccuracy of the GT label map with respect to the OM data, leading to a possible inaccuracy of the defect location of around $380\,\mu\mathrm{m}$. By manually refining the registration, the inaccuracy can be further reduced. Nevertheless, it has to be considered when analyzing the network performance, as the registration error can propagate through the CNN training. This might lead to a systematic error in defect localization. For the given use case, this error seems acceptable as the exact localization of the defect is of secondary importance to the pure detection of the defect.

7 Conclusion & Outlook

The data science pipeline developed in this thesis is able to automatically detect pores in titanium AM parts based on process monitoring data. This is an important step for the industrial implementation of metal AM in safety-critical applications as it allows for an in-process evaluation of the part quality. The in-process evaluation drastically reduces the time and cost for quality assurance by reducing post-process NDT steps. As the NDT of AM parts makes up to 50% of the part costs, reducing those costs will make AM more competitive compared to conventional manufacturing methods. Additionally, the developed evaluation system is not limited by the part size (in contrast to CT). Hence, it enables the evaluation and manufacturing of large AM components that are too big for conventional CT inspection.

Within the developed system, the training of CNNs to automatically analyze and segment NDT data (CT and OM) shows great potential. Firstly, the state-of-the-art U-Net architecture is adapted from medical image analysis and trained iteratively to segment pores in CT data. The iterative refinement of the ground truth label map by human interaction leads to a well-performing model with reasonable effort. The additional revision of selected label maps by qualified CT inspectors creates a qualified test dataset. This qualified label map allows for the quantification of the model performance. The quantification is based on the developed use-case-specific metrics. Besides adopting the POD to deep learning, the metrics BUD, NUD_{400} and FP_{400} are introduced to evaluate the model regarding the defects relevant to the investigated use case. For the CT CNN, a POD of around $330\,\mu m$ is achieved. The CT CNN is used to create the ground truth training data as a reference for the training of the OM CNN. By combining the information about the defect size and location from the CT CNN with the online monitoring data, a supervised training of the OM CNN is feasible. This allows for the interpretation of the online monitoring data with respect to actual existing pores and not solely with regard to process deviations (which might not result in pores).

The feasibility of this approach is shown by training a selected U-Net architecture. The quantitative and qualitative evaluation of the model performance shows that the U-Net can be successfully trained to detect pores in metal AM parts based on the online monitoring data. Specific models can detect all pores in the qualified dataset larger than $531\,\mu m$. This comes at the cost of an increased false positive count. The selection of suitable models is, therefore, use-case-dependent. An increased understanding of the model behavior and its limitations is achieved through additional investigations. This should facilitate the qualification of the developed system for industrial use.

In summary, the presented work lays the baseline for the automatic evaluation of online monitoring data of the LPBF process by CNNs. It successfully combines NDT, AM, and AI to improve the non-destructive testing of metal-printed parts and shows that pores can be reliably detected in test specimens.

Based on these findings and the current limitations outlined in Section 6.3, further research steps will be discussed in the following. The current approach provides a blueprint for further development. The developed pipeline can incorporate additional sensors to improve the performance of the overall system. The current photodiodes could be supplemented by, e.g., acoustic sensors or a thermal camera. The additional sensors might allow for the detection of other defect mechanisms or increase the performance of already investigated defects. For example, the formation of cracks in lower regions of the part cannot be reliably detected by the photodiodes as they might not influence the meltpool at higher levels. Such defect mechanisms might, for example, be detected by an acoustic emission sensor.

Furthermore, the transfer to other printers will provide additional confidence in the developed algorithm. The current findings are based on a continuous laser machine with titanium powder. Ongoing research in a follow-up project investigates the transferability of the trained network to a second printer with a pulsed laser. For this transfer, the pre-processing steps for the volume generation will be adapted to incorporate the pulsed laser. This should allow for an optimal input to the CNN. The adaptation to new printers should be facilitated by using the originally trained CNN as a basis and applying a transfer learning approach.

Besides the transfer to other printers, the transferability of the network to other defect types, such as inclusions, will also be assessed in this study. Inclusions have a higher density than the specimen material, in contrast to pores, which have a lower density than the specimen material. Therefore, they might result in a different meltpool signature and, subsequently, a different photodiode signal than pores. If this is the case for different defect types, it might be necessary to train the CNN for each defect type individually. A numerical simulation of different defects in the meltpool might provide further insight.

Additionally, a long-term study accompanying the potential introduction of the system to safety-critical applications should evaluate the performance of the model on real-life defects. Creating "real" defects in a research scope is not feasible as this would require a vast amount of printed parts due to the high print quality. To evaluate the model on "real" defects, the model should first be used in parallel with the conventional CT inspection. The quality assurance would be performed by CT evaluation, but the model would already be used to analyze the part quality independently. By gathering long-term data and comparing the model performance to the reference CT scans, a high degree of trust can be established. This allows for the generation of high-quality ground truth data upon which the model can be evaluated. Once sufficient data and trust are gathered, a successive transfer from CT to OM might be performed.

The introduction of the system to industry also requires defined qualification standards. For the aviation sector, the EASA is currently developing high-level guidelines for the qualification of AI [EASA, 2020, 2021]. Those guidelines should be adapted to AI NDT and combined with well-established NDT metrics such as the POD. Further research should be conducted with regard to the statistical methods used to quantify the performance of AI applications. In particular, different methods for the POD calculation should be reviewed and adapted.

Once the system is implemented for inline quality control, the next possible step is the development of a feedback control. This feedback control takes the information of the monitoring system as input and adjusts the printing process accordingly. The easiest feedback control would be to stop the printer once a defect is detected. This way, resources (metal powder and print time) can be saved. The more complex feedback control would enable the printer to "repair" defective regions, e.g., by remelting such regions. This would also require a better understanding of the melting process and the monitored phenomena. A numerical simulation of the meltpool, the effect of defects on it, and, in particular, the associated meltpool radiation might provide further insights.

Overall, the developed online monitoring solution is a first step towards the automatic in-process non-destructive testing of additive manufactured parts. In contrast to previous studies, the approach focuses on actual defects as detected by CT and not only on deviations in the process that might not lead to actual defects. Therefore, it tackles the two most significant challenges identified by industry and the World Economic Forum: cost reduction and quality assurance [Basso et al., 2022]. Firstly, reducing or replacing the cost-intensive CT inspection will significantly decrease the costs of AM parts, particularly the NDT costs. Secondly, it will allow for the production and testing of large-scale metal AM parts for safety-critical applications and, therefore, acts as an enabler for the production of large components in the aerospace industry.

References

Alkadhi, Hatem, Leschka, Sebastian, Stolzmann, Paul, and Scheffel, Hans. *Wie funktioniert CT? Eine Einführung in Physik, Funktionsweise und klinische Anwendungen der Computertomographie*. Springer-Verlag GmbH Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 9783642178023. doi: 10.1007/978-3-642-17803-0. URL http://dx.doi.org/10.1007/978-3-642-17803-0.

Antonelli, Michela, Reinke, Annika, Bakas, Spyridon, Farahani, Keyvan, AnnetteKopp-Schneider, Landman, Bennett A., Litjens, Geert, Menze, Bjoern, Ronneberger, Olaf, Summers, Ronald M., van Ginneken, Bram, Bilello, Michel, Bilic, Patrick, Christ, Patrick F., Do, Richard K. G., Gollub, Marc J., Heckers, Stephan H., Huisman, Henkjan, Jarnagin, William R., McHugo, Maureen K., Napel, Sandy, Pernicka, Jennifer S. Goli, Rhode, Kawal, Tobon-Gomez, Catalina, Vorontsov, Eugene, Meakin, James A., Ourselin, Sebastien, Wiesenfarth, Manuel, Arbelaez, Pablo, Bae, Byeonguk, Chen, Sihong, Daza, Laura, Feng, Jianjiang, He, Baochun, Isensee, Fabian, Ji, Yuanfeng, Jia, Fucang, Kim, Namkug, Kim, Ildoo, Merhof, Dorit, Pai, Akshay, Park, Beomhee, Perslev, Mathias, Rezaiifar, Ramin, Rippel, Oliver, Sarasua, Ignacio, Shen, Wei, Son, Jaemin, Wachinger, Christian, Wang, Liansheng, Wang, Yan, Xia, Yingda, Xu, Daguang, Xu, Zhanwei, Zheng, Yefeng, Simpson, Amber L., Maier-Hein, Lena, and Cardoso, M. Jorge. The medical segmentation decathlon. 2021. URL http://arxiv.org/pdf/2106.05735v1.

Arridge, S. R. Optical tomography in medical imaging. *Inverse Problems*, 15(2):R41–R93, 1999. ISSN 0266-5611. doi: 10.1088/0266-5611/15/2/022.

Bamberg, Joachim, Zenzinger, Günter, and Ladewig, Alexander. In-process control of selective laser melting by quantitative optical tomography. 2016.

Bartlett, Jamison L., Heim, Frederick M., Murty, Yellapu V., and Li, Xiaodong. In situ defect detection in selective laser melting via full-field infrared thermography. *Additive Manufacturing*, 24:595–605, 2018. ISSN 22148604. doi: 10.1016/j.addma.2018.10.045.

Basso, Maria, Betti, Francisco, Cronin, Ian, and Schönfuß, Benjamin. An additive manufacturing breakthrough: A how-to guide for scaling and overcoming key challenges. 2022, 2022.

Baumgartl, Hermann, Tomas, Josef, Buettner, Ricardo, and Merkel, Markus. A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring: Burg, stadt, residenz ; ein kurzer abriss seiner geschichte. *Progress in Additive Manufacturing*, 96(5):216, 2020. ISSN 2363-9512. doi: 10.1007/s40964-019-00108-3.

Berens, A. P. and Hovey, P. W. Flaw detection reliability criteria: Volume 1 - methods and results. 1984.

Berumen, Sebastian, Bechmann, Florian, Lindner, Stefan, Kruth, Jean-Pierre, and Craeghs, Tom. Quality control of laser- and powder bed-based additive manufacturing (am) technologies. *Physics Procedia*, 5:617–622, 2010. ISSN 18753892. doi: 10.1016/j.phpro.2010.08.089.

Betts, Scott B. The relationship between in-process quality metrics & computational tomography. 2019.

Betts, Scott B. and Cola, Mark. Evaluation of quality signatures™ using in-situ process control during additive manufacturing with aluminum alloy alsi10mg – part 2. 2018a.

Betts, Scott B. and Cola, Mark. In-situ process mapping using thermal quality signatures™ during additive manufacturing with titanium alloy ti-6al-4v. 2018b.

Betts, Scott B. and Jacquemetton, Lars. In-situ melt pool "thermal signature" defect detection of recoater failure using co-axial planck thermometry. 2018.

Betts, Scott B. and Jacquemetton, Lars. Support structure optimization using printrite3d. 2019.

Bidare, P., Bitharas, I., Ward, R. M., Attallah, M. M., and Moore, A. J. Fluid and particle dynamics in laser powder bed fusion. *Acta Materialia*, 142:107–120, 2018. ISSN 13596454. doi: 10.1016/j.actamat.2017.09.051.

Bohm, Arno, Kielanowski, Piotr, and Mainland, G. Bruce. *Quantum Physics*. Springer Netherlands, Dordrecht, 2019. ISBN 978-94-024-1758-6. doi: 10.1007/978-94-024-1760-9.

Brailovski, Vladimir, Kalinicheva, Victoria, Letenneur, Morgan, Lukashevich, Konstantin, Sheremetyev, Vadim, and Prokoshkin, Sergey. Control of density and grain structure of a laser powder bed-fused superelastic ti-18zr-14nb alloy: Simulation-driven process mapping. *Metals*, 10(12):1697, 2020. doi: 10.3390/met10121697.

British Standards Institution. Additive manufacturing — general principles — terminology, 2015.

Brunel, Nicolas, Hakim, Vincent, and Richardson, Magnus J. E. Single neuron dynamics and computation. *Current opinion in neurobiology*, 25:149–155, 2014. doi: 10.1016/j.conb.2014.01.005.

Buduma, Nikhil and Locascio, Nicholas. *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. First edition edition, 2017. ISBN 978-1491925614.

Buzug, Thorsten M. *Computed Tomography*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-39407-5. doi: 10.1007/978-3-540-39408-2.

Calignano, Flaviana, editor. *Additive Manufacturing (AM) of Metallic Alloys*. MDPI - Multidisciplinary Digital Publishing Institute, Basel, Switzerland, 2020. ISBN 9783039431403. URL https://directory.doabooks.org/handle/20.500.12854/69093.

Carl, Volker. Monitoring system for the quality assessment in additive manufacturing. 2015.

Carmignato, Simone, Dewulf, Wim, and Leach, Richard. *Industrial X-Ray Computed Tomography*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-59571-9. doi: 10.1007/978-3-319-59573-3.

Casati, R., Lemke, J., and Vedani, M. Microstructure and fracture behavior of 316l austenitic stainless steel produced by selective laser melting. *Journal of Materials Science & Technology*, 32(8):738–744, 2016. ISSN 10050302. doi: 10.1016/j.jmst.2016.06.016.

Cheng, R.C.H. and Iles, T. C. One-sided confidence bands for cumulative distribution functions. 1988.

Chollet, François. *Deep learning with Python*. Safari Tech Books Online. Manning, Shelter Island, NY, 2018. ISBN 978-1-61729-443-3.

Cicek, Özgün, Abdulkadir, Ahmed, Lienkamp, Soeren S., Brox, Thomas, and Ronneberger, Olaf. 3d u-net: Learning dense volumetric segmentation from sparse annotation. 2016.

Cierniak, Robert. *X-Ray Computed Tomography in Biomedical Engineering*. Springer London, London, 2011. ISBN 978-0-85729-026-7. doi: 10.1007/978-0-85729-027-4.

Clijsters, S., Craeghs, T., Buls, S., Kempen, K., and Kruth, J.-P. In situ quality control of the selective laser melting process using a high-speed, real-time melt pool monitoring system. *The International Journal of Advanced Manufacturing Technology*, 75(5-8):1089–1101, 2014. ISSN 0268-3768. doi: 10.1007/s00170-014-6214-8.

Cox, Robert W., Ashburner, John, Breman, Hester, Fissell, Kate, Haselgrove, Christian, Holmes, Colin J., Lancaster, Jack L., Rex, David E., Smith, Stephen M., Woodward, Jeffery B., and Strother, Stephen C. A (sort of) new image data format standard: Nifti-1. 2004.

Craeghs, Tom, Bechmann, Florian, Berumen, Sebastian, and Kruth, Jean-Pierre. Feedback control of layerwise laser melting using optical sensors. *Physics Procedia*, 5:505–514, 2010. ISSN 18753892. doi: 10.1016/j.phpro.2010.08.078.

Craeghs, Tom, Clijsters, Stijn, Kruth, Jean.-Pierre, Bechmann, Florian, and Ebert, Marie.-Christin. Detection of process failures in layerwise laser melting with optical process monitoring. *Physics Procedia*, 39: 753–759, 2012. ISSN 18753892. doi: 10.1016/j.phpro.2012.10.097.

Crum, William R., Camara, Oscar, and Hill, Derek L. G. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11):1451–1461, 2006. doi: 10.1109/TMI.2006.880587.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. ISSN 0932-4194. doi: 10.1007/BF02551274.

Daugaard Jørgensen, Mia, Antulov, Ronald, Hess, Søren, and Lysdahlgaard, Simon. Convolutional neural network performance compared to radiologists in detecting intracranial hemorrhage from brain computed tomography: A systematic review and meta-analysis. *European journal of radiology*, 146:110073, 2022. doi: 10.1016/j.ejrad.2021.110073.

Dice, Lee R. Measures of the amount of ecologic association between species. 1945.

Diehl, Brett, Castro, Alberto, Jaquemetton, Lars, and Beckett, Darren. Technical focus: In situ monitoring for additive manufacturing. *Materials Evaluation*, 80(4):64–73, 2022. ISSN 00255327. doi: 10.32548/2022.me-04271.

Dillhoefer, Alexander, Rieder, Hans, Spies, Martin, Bamberg, Joachim, and Hess, Thomas. Online monitoring of additive manufacturing processes using ultrasound. 2014.

DIN. Schweißen und verwandte prozesse – einteilung von geometrischen unregelmäßigkeiten an metallischen werkstoffen: Teil 1: Schmelzschweißen (iso 6520-1:2007); dreisprachige fassung en iso 6520-1:2007, 2007.

Dong, Zhichao, Liu, Yabo, Wen, Weibin, Ge, Jingran, and Liang, Jun. Effect of hatch spacing on melt pool and as-built quality during selective laser melting of stainless steel: Modeling and experimental approaches. *Materials (Basel, Switzerland)*, 12(1), 2018. ISSN 1996-1944. doi: 10.3390/ma12010050.

Doubenskaia, M. Comprehensive optical monitoring of selective laser melting. *Journal of Laser Micro/Nanoengineering*, 7(3):236–243, 2012. doi: 10.2961/jlmn.2012.03.0001.

Doubenskaia, Maria, Grigoriev, Sergey, Zhirnov, Ivan, and Smurov, Igor. Parametric analysis of slm using comprehensive optical monitoring. *Rapid Prototyping Journal*, 22(1):40–50, 2016. ISSN 1355-2546. doi: 10.1108/RPJ-04-2014-0046.

Du, Ke-Lin and Swamy, M. N. S. *Neural Networks and Statistical Learning*. Springer London, London, 2019. ISBN 978-1-4471-7451-6. doi: 10.1007/978-1-4471-7452-3.

Duong, Emil, Masseling, Lukas, Knaak, Christian, Dionne, Paul, and Megahed, Mustafa. Scan path resolved thermal modelling of lpbf. *Additive Manufacturing Letters*, 3:100047, 2022. ISSN 27723690. doi: 10.1016/j.addlet.2022.100047.

EASA. Artificial intelligence roadmap: A human-centric approach to ai in aviation. 2020.

EASA. Easa concept paper: First usable guidance for level 1 machine learning applications: A deliverable of the easa ai roadmap. 2021.

Eschner, N., Weiser, L., Häfner, B., and Lanza, G. Classification of specimen density in laser powder bed fusion (l-pbf) using in-process structure-borne acoustic process emissions. *Additive Manufacturing*, 34: 101324, 2020. ISSN 22148604. doi: 10.1016/j.addma.2020.101324.

Everton, Sarah K., Hirsch, Matthias, Stravroulakis, Petros, Leach, Richard K., and Clare, Adam T. Review of in-situ process monitoring and in-situ metrology for metal additive manufacturing. *Materials & Design*, 95:431–445, 2016. ISSN 02641275. doi: 10.1016/j.matdes.2016.01.099.

Falk, Thorsten, Mai, Dominic, Bensch, Robert, Çiçek, Özgün, Abdulkadir, Ahmed, Marrakchi, Yassine, Böhm, Anton, Deubner, Jan, Jäckel, Zoe, Seiwald, Katharina, Dovzhenko, Alexander, Tietz, Olaf, Dal Bosco, Cristina, Walsh, Sean, Saltukoglu, Deniz, Tay, Tuan Leng, Prinz, Marco, Palme, Klaus, Simons, Matias, Diester, Ilka, Brox, Thomas, and Ronneberger, Olaf. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. doi: 10.1038/s41592-018-0261-2.

Feldkamp, L. A., Davis, L. C., and Kress, J. W. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612, 1984. ISSN 1084-7529. doi: 10.1364/JOSAA.1.000612.

Ferrar, B., Mullen, L., Jones, E., Stamp, R., and Sutcliffe, C. J. Gas flow effects on selective laser melting (slm) manufacturing performance. *Journal of Materials Processing Technology*, 212(2):355–364, 2012. ISSN 09240136. doi: 10.1016/j.jmatprotec.2011.09.020.

Fischer, Felix Gabriel, Zimmermann, Max Gero, Praetzsch, Niklas, and Knaak, Christian. Monitoring of the powder bed quality in metal additive manufacturing using deep transfer learning. *Materials & Design*, 222:111029, 2022. ISSN 02641275. doi: 10.1016/j.matdes.2022.111029.

Frye, Roger, Xuan Yu, Christina, Betts, Scott B., Jacquemetton, Lars, and Anderson, Kevin C. Printrite3d machine learning case study. 2020.

Fuchs, Lukas and Eischer, Christopher. In-process monitoring systems for metal additive manufacturing. 2018, 2018.

Gobert, Christian, Reutzel, Edward W., Petrich, Jan, Nassar, Abdalla R., and Phoha, Shashi. Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. *Additive Manufacturing*, 21:517–528, 2018. ISSN 22148604. doi: 10.1016/j.addma.2018.04.005.

Goegelein, A., Ladewig, A., Zenzinger, G., and Bamberg, J. Process monitoring of additive manufacturing by using optical tomography. 2018.

Goh, G. D., Sing, S. L., and Yeong, W. Y. A review on machine learning in 3d printing: applications, potential, and challenges. *Artificial Intelligence Review*, 2020. ISSN 0269-2821. doi: 10.1007/s10462-020-09876-9.

Gong, Haijun, Rafi, Khalid, Gu, Hengfeng, Starr, Thomas, and Stucker, Brent. Analysis of defect generation in ti–6al–4v parts made using powder bed fusion additive manufacturing processes. *Additive Manufacturing*, 1-4:87–98, 2014. ISSN 22148604. doi: 10.1016/j.addma.2014.08.002.

Gong, Haijun, Rafi, Khalid, Gu, Hengfeng, Janaki Ram, G. D., Starr, Thomas, and Stucker, Brent. Influence of defects on mechanical properties of ti–6al–4v components produced by selective laser melting and electron beam melting. *Materials & Design*, 86:545–554, 2015. ISSN 02641275. doi: 10.1016/j.matdes.2015.07.147.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. *Deep Learning*. MIT Press, 2016. URL `http://www.deeplearningbook.org/`.

Gordon, Jerard V., Narra, Sneha P., Cunningham, Ross W., Liu, He, Chen, Hangman, Suter, Robert M., Beuth, Jack L., and Rollett, Anthony D. Defect structure process maps for laser powder bed fusion additive manufacturing. *Additive Manufacturing*, 36:101552, 2020. ISSN 22148604. doi: 10.1016/j.addma.2020.101552.

Grasso, Marco and Colosimo, Bianca Maria. Process defects and in situ monitoring methods in metal powder bed fusion: a review. *Measurement Science and Technology*, 28(4), 2017. doi: 10.1088/1361-6501/aa5c4f.

Guo, Weihong "Grace", Tian, Qi, Guo, Shenghan, and Guo, Yuebin. A physics-driven deep learning model for process-porosity causal relationship and porosity prediction with interpretability in laser metal deposition. *CIRP Annals*, 69(1):205–208, 2020. ISSN 00078506. doi: 10.1016/j.cirp.2020.04.049.

Gurney, Kevin. *An Introduction to Neural Networks*. Taylor and Francis, Hoboken, 2003. ISBN 9781857286731. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=182103.

Härdle, Wolfgang Karl, Klinke, Sigbert, and Rönz, Bernd. *Introduction to Statistics*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-17703-8. doi: 10.1007/978-3-319-17704-5.

Harrison, Neil J., Todd, Iain, and Mumtaz, Kamran. Reduction of micro-cracking in nickel superalloys processed by selective laser melting: A fundamental alloy design approach. *Acta Materialia*, 94:59–68, 2015. ISSN 13596454. doi: 10.1016/j.actamat.2015.04.035.

Hertel, Ingolf V. and Schulz, C.-P. *Atome, Moleküle und optische Physik 1*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-53103-7. doi: 10.1007/978-3-662-53104-4.

Holtmann, Jonas, Kiefel, Denis, Neumann, Stefan, Stoessel, Rainer, and Grosse, Christian U. A data driven approach to the online monitoring of the additive manufacturing process. *Advanced Materials Research*, 1161:137–144, 2021a. doi: 10.4028/www.scientific.net/AMR.1161.137.

Holtmann, Jonas, Permadi, Jordy, Kiefel, Denis, and Grosse, Christian. Convolutional neural networks for the automated segmentation of computed tomography scans. *Special Issue of e-Journal of Nondestructive Testing (eJNDT)*, (Vol. 27(6)), 2021b.

Hovey, P. W. and Berens, A. P. Statistical evaluation of nde reliability in the aerospace industry. 1988.

Hu, Y. N., Wu, S. C., Withers, P. J., Zhang, J., Bao, H.Y.X., Fu, Y. N., and Kang, G. Z. The effect of manufacturing defects on the fatigue life of selective laser melted ti-6al-4v structures. *Materials & Design*, 192:108708, 2020. ISSN 02641275. doi: 10.1016/j.matdes.2020.108708.

Huang, Yuze, Fleming, Tristan G., Clark, Samuel J., Marussi, Sebastian, Fezzaa, Kamel, Thiyagalingam, Jeyan, Leung, Chu Lun Alex, and Lee, Peter D. Keyhole fluctuation and pore formation mechanisms during laser powder bed fusion additive manufacturing. *Nature communications*, 13(1):1170, 2022. doi: 10.1038/s41467-022-28694-x.

Hügel, Helmut and Graf, Thomas. *Laser in der Fertigung: Strahlquellen, Systeme, Fertigungsverfahren*. Aus dem Programm Fertigung. Vieweg + Teubner, Wiesbaden, 2., neu bearb. aufl. edition, 2009. ISBN 383510005X.

Isensee, Fabian, Jaeger, Paul F., Kohl, Simon A. A., Petersen, Jens, and Maier-Hein, Klaus H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18 (2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

ISO. Additive manufacturing — design — part 1: Laser-based powder bed fusion of metals, 07.2019.

Johnson, Hans J., McCormick, Matthew M., and Ibanez, Luis. *The Itk Software Guide Book 2: Design and Functionality*. Kitware, Incorporated, 2015. ISBN 9781930934283.

Kaehn, Raffael. *Digital Image Processing for In-Situ Defect Detection in Laser Powder Bed Fusion*. Bachelor thesis, Hochschule Bremen, 2020.

Kerfoot, Eric, Clough, James, Oksuz, Ilkay, Lee, Jack, King, Andrew P., and Schnabel, Julia A. Left-ventricle quantification using residual u-net. In Pop, Mihaela, Sermesant, Maxime, Zhao, Jichao, Li, Shuo, McLeod, Kristin, Young, Alistair, Rhode, Kawal, and Mansi, Tommaso, editors, *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, volume 11395 of *Lecture Notes in Computer Science*, pages 371–380. Springer International Publishing, Cham, 2019. ISBN 978-3-030-12028-3. doi: 10.1007/978-3-030-12029-0{\textunderscore}40.

Khorasani, AmirMahyar, Gibson, Ian, Awan, Umar Shafique, and Ghaderi, Alireza. The effect of slm process parameters on density, hardness, tensile strength and surface quality of ti-6al-4v. 2019.

Kiefel, Denis. *Quantitative Porositätscharakterisierung von CFK-Werkstoffen mit der Mikro-Computertomografie*. Dissertation, Technische Universität München, 2017.

King, Wayne E., Barth, Holly D., Castillo, Victor M., Gallegos, Gilbert F., Gibbs, John W., Hahn, Douglas E., Kamath, Chandrika, and Rubenchik, Alexander M. Observation of keyhole-mode laser melting in laser powder-bed fusion additive manufacturing. *Journal of Materials Processing Technology*, 214(12):2915–2925, 2014. ISSN 09240136. doi: 10.1016/j.jmatprotec.2014.06.005.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. 2017. doi: 10.48550/arXiv.1412.6980. URL `http://arxiv.org/pdf/1412.6980v9`.

Kratz, Bärbel. *Reduktion von Metallartefakten in der Computertomographie*. Springer Fachmedien Wiesbaden, Wiesbaden, 2015. ISBN 978-3-658-08420-2. doi: 10.1007/978-3-658-08421-9.

Krieger, Hanno. *Grundlagen der Strahlungsphysik und des Strahlenschutzes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-55759-4. doi: 10.1007/978-3-662-55760-0.

Lachmayer, Roland and Lippert, Rene Bastian. *Additive Manufacturing Quantifiziert*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. ISBN 978-3-662-54112-8. doi: 10.1007/978-3-662-54113-5.

Ladewig, Alexander. *Optische Tomographie: Online Prozessüberwachung für das selektive Laserschmelzen*. Dissertation, Karlsruher Institut für Technologie, Karlsruhe, 2019.

Ladewig, Alexander, Schlick, Georg, Fisser, Maximilian, Schulze, Volker, and Glatzel, Uwe. Influence of the shielding gas flow on the removal of process by-products in the selective laser melting process. *Additive Manufacturing*, 10:1–9, 2016. ISSN 22148604. doi: 10.1016/j.addma.2016.01.004.

Lane, Brandon, Zhirnov, Ivan, Mekhontsev, Sergey, Grantham, Steven, Ricker, Richard, Rauniyar, Santosh, and Chou, Kevin. Transient laser energy absorption, co-axial melt pool monitoring, and relationship to melt pool morphology. *Additive Manufacturing*, 36, 2020. ISSN 22148604. doi: 10.1016/j.addma.2020.101504.

Larobina, Michele and Murino, Loredana. Medical image file formats. *Journal of digital imaging*, 27(2): 200–206, 2014. doi: 10.1007/s10278-013-9657-9.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. 1989.

LeCun, Y., Bottou, Leon, Orr, Genevieve B., and Müller, Klaus-Robert. Efficient backprop. 1998.

Letenneur, Morgan, Kreitcberg, Alena, and Brailovski, Vladimir. Optimization of laser powder bed fusion processing using a combination of melt pool modeling and design of experiment approaches: Density control. *Journal of Manufacturing and Materials Processing*, 3(1):21, 2019. doi: 10.3390/jmmp3010021.

Li, E. L., Wang, L., Yu, A. B., and Zhou, Z. Y. A three-phase model for simulation of heat transfer and melt pool behaviour in laser powder bed fusion process. *Powder Technology*, 381:298–312, 2021. ISSN 00325910. doi: 10.1016/j.powtec.2020.11.061.

Li, Xiangrui, Morgan, Paul S., Ashburner, John, Smith, Jolinda, and Rorden, Christopher. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of neuroscience methods*, 264:47–56, 2016. doi: 10.1016/j.jneumeth.2016.03.001.

Listl, Marek and Orye, Davy. Correlation of in-process monitoring data and mechanical properties of lattice structures: In-situ process monitoring of lattice structures manufactured by dmls®. 2019.

Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, van der Laak, Jeroen A. W. M., van Ginneken, Bram, and Sánchez, Clara I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(13):60–88, 2017. ISSN 13618415. doi: 10.1016/j.media.2017.07.005. URL http://arxiv.org/pdf/1702.05747v2.

Liu, Qian Chu, Elambasseril, Joe, Sun, Shou Jin, Leary, Martin, Brandt, Milan, and Sharp, Peter Khan. The effect of manufacturing defects on the fatigue behaviour of ti-6al-4v specimens fabricated using selective laser melting. *Advanced Materials Research*, 891-892:1519–1524, 2014. doi: 10.4028/www.scientific.net/AMR.891-892.1519.

Liu, Xiaoxuan, Faes, Livia, Kale, Aditya U., Wagner, Siegfried K., Fu, Dun Jack, Bruynseels, Alice, Mahendiran, Thushika, Moraes, Gabriella, Shamdas, Mohith, Kern, Christoph, Ledsam, Joseph R., Schmid, Martin K., Balaskas, Konstantinos, Topol, Eric J., Bachmann, Lucas M., Keane, Pearse A., and Denniston, Alastair K. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019. ISSN 25897500. doi: 10.1016/S2589-7500(19)30123-2.

London, Michael and Häusser, Michael. Dendritic computation. *Annual review of neuroscience*, 28:503–532, 2005. ISSN 0147-006X. doi: 10.1146/annurev.neuro.28.061604.135703.

Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. 2015. URL http://arxiv.org/pdf/1411.4038v2.

Lowekamp, Bradley C., Chen, David T., Ibáñez, Luis, and Blezek, Daniel. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013. ISSN 1662-5196. doi: 10.3389/fninf.2013.00045.

Lu, Xufei, Cervera, Miguel, Chiumenti, Michele, and Lin, Xin. Residual stresses control in additive manufacturing. *Journal of Manufacturing and Materials Processing*, 5(4):138, 2021. doi: 10.3390/jmmp5040138.

Mahmoud, Dalia, Magolon, Marcin, Boer, Jan, Elbestawi, M. A., and Mohammadi, Mohammad Ghayoomi. Applications of machine learning in process monitoring and controls of l-pbf additive manufacturing: A review. *Applied Sciences*, 11(24):11910, 2021. doi: 10.3390/app112411910.

Mancisidor, A. M., Garciandia, F., Sebastian, M. San, Álvarez, P., Díaz, J., and Unanue, I. Reduction of the residual porosity in parts manufactured by selective laser melting using skywriting and high focus offset strategies. *Physics Procedia*, 83:864–873, 2016. ISSN 18753892. doi: 10.1016/j.phpro.2016.08.090.

Martin, Aiden A., Calta, Nicholas P., Hammons, Joshua A., Khairallah, Saad A., Nielsen, Michael H., Shuttlesworth, Richard M., Sinclair, Nicholas, Matthews, Manyalibo J., Jeffries, Jason R., Willey, Trevor M., and Lee, Jonathan R.I. Ultrafast dynamics of laser-metal interactions in additive manufacturing alloys

captured by in situ x-ray imaging. *Materials Today Advances*, 1:100002, 2019a. ISSN 25900498. doi: 10.1016/j.mtadv.2019.01.001.

Martin, Aiden A., Calta, Nicholas P., Khairallah, Saad A., Wang, Jenny, DePond, Phillip J., Fong, Anthony Y., Thampy, Vivek, Guss, Gabe M., Kiss, Andrew M., Stone, Kevin H., Tassone, Christopher J., Nelson Weker, Johanna, Toney, Michael F., van Buuren, Tony, and Matthews, Manyalibo J. Dynamics of pore formation during laser powder bed fusion additive manufacturing. *Nature communications*, 10(1): 1987, 2019b. doi: 10.1038/s41467-019-10009-2.

Mattes, David, Haynor, David R., Vesselle, Hubert, Lewellen, Thomas K., and Eubank, William. Pet-ct image registration in the chest using free-form deformations. *IEEE Transactions on Medical Imaging*, 22 (1):120–128, 2003. ISSN 0278-0062. doi: 10.1109/TMI.2003.809072.

McCann, Ronan, Obeidi, Muhannad A., Hughes, Cian, McCarthy, Éanna, Egan, Darragh S., Vijayaraghavan, Rajani K., Joshi, Ajey M., Acinas Garzon, Victor, Dowling, Denis P., McNally, Patrick J., and Brabazon, Dermot. In-situ sensing, process monitoring and machine control in laser powder bed fusion: A review. *Additive Manufacturing*, 45:102058, 2021. ISSN 22148604. doi: 10.1016/j.addma.2021.102058.

McGowan, Erin, Gawade, Vidita, and Guo, Weihong Grace. A physics-informed convolutional neural network with custom loss functions for porosity prediction in laser metal deposition. *Sensors (Basel, Switzerland)*, 22(2), 2022. doi: 10.3390/s22020494.

Meboldt, Mirko and Klahn, Christoph. *Industrializing Additive Manufacturing - Proceedings of Additive Manufacturing in Products and Applications - AMPA2017*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-66865-9. doi: 10.1007/978-3-319-66866-6.

Megahed, Mustafa, Mindt, Hans-Wilfried, Willems, Jöerg, Dionne, Paul, Jacquemetton, Lars, Craig, James, Ranade, Piyush, and Peralta, Alonso. Lpbf right the first time—the right mix between modeling and experiments. 2019. doi: 10.1007/s40192-019-00133-8.

Meiners, Wilhelm. *Direktes selektives Laser Sintern einkomponentiger metallischer Werkstoffe*. Dissertation, RWTH Aachen, Aachen, 1999.

Meng, Lingbin, McWilliams, Brandon, Jarosinski, William, Park, Hye-Yeong, Jung, Yeon-Gil, Lee, Jehyun, and Zhang, Jing. Machine learning in additive manufacturing: A review. *JOM*, 72(6):2363–2377, 2020. ISSN 1047-4838. doi: 10.1007/s11837-020-04155-y.

Mercelis, Peter and Kruth, Jean-Pierre. Residual stresses in selective laser sintering and selective laser melting. *Rapid Prototyping Journal*, 12(5):254–265, 2006. ISSN 1355-2546. doi: 10.1108/13552540610707013.

Milcke, Björn. *Explainability of Machine Learning Algorithms in Quality Assurance Based on the Example of Inline Monitoring for Additive Manufacturing*. Master thesis, Universität Bremen, Bremen, 2022.

Milletari, Fausto, Navab, Nassir, and Ahmadi, Seyed-Ahmad. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016.

Mohr, Gunther, Altenburg, Simon J., Ulbricht, Alexander, Heinrich, Philipp, Baum, Daniel, Maierhofer, Christiane, and Hilgenberg, Kai. In-situ defect detection in laser powder bed fusion by using thermography and optical tomography—comparison to computed tomography. *Metals*, 10(1):103, 2020. doi: 10.3390/met10010103.

Montijo Badeira, Paulo. *Data Registration of Computed Tomography and Online Monitoring Data in Additive Manufacturing*. Master thesis, Technische Universität München, 2020.

Moon, Seunghyun, Ma, Ruimin, Attardo, Ross, Tomonto, Charles, Nordin, Mark, Wheelock, Paul, Glavicic, Michael, Layman, Maxwell, Billo, Richard, and Luo, Tengfei. Impact of surface and pore characteristics on fatigue life of laser powder bed fusion ti-6al-4v alloy described by neural network models. *Scientific reports*, 11(1):20424, 2021. doi: 10.1038/s41598-021-99959-6.

Nic Ma, Wenqi Li, Richard Brown, Yiheng Wang, Behrooz, Benjamin Gorman, Hans Johnson, Isaac Yang, Eric Kerfoot, charliebudd, Mohammad Adil, Yiwen Li, Yuan-Ting Hsieh, Arpit Aggarwal, masadcv, Cameron Trentz, adam aji, myron, Mark Graham, Ben Murray, Gagan Daroach, Petru-Daniel Tudosiu, Matt McCormick, Ali Hatamizadeh, Ambros, Balamurali, Christian Baker, Holger Roth, and Jan Sellner. Project-monai/monai: 0.6.0, 2021.

Oliveira, J. P., LaLonde, A. D., and Ma, J. Processing parameters in laser powder bed fusion metal additive manufacturing. *Materials & Design*, 193:108762, 2020. ISSN 02641275. doi: 10.1016/j.matdes.2020.108762.

Otsu, Nobuyuki. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. ISSN 0018-9472. doi: 10.1109/TSMC.1979.4310076.

Paszke, Adam, Gross, Sam, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. 2019.

Permadi, Jordy. *Automated Anomaly Detection in Computed Tomography Scan Images of Additive Manufactured Titanium Components using Deep Learning*. Master thesis, Technische Universität München, München, 2021.

Platacis, Ernests, Kaldre, Imants, Blumbergs, Ervīns, Goldšteins, Linards, and Serga, Vera. Titanium production by magnesium thermal reduction in the electroslag process. *Scientific reports*, 9(1):17566, 2019. doi: 10.1038/s41598-019-54112-2.

Podgoršak, E. B. *Radiation physics for medical physicists*. Biological and medical physics, biomedical engineering. Springer, Heidelberg, 2nd, enl. ed. edition, 2010. ISBN 9783642008740.

Poprawe, Reinhart. *Lasertechnik für die Fertigung: Grundlagen, Perspektiven und Beispiele für den innovativen Ingenieur ; mit 26 Tabellen*. VDI-Buch. Springer, Berlin and Heidelberg, 2005. ISBN 9783540214069.

Quintino, L. and Assunção, E. Conduction laser welding. In *Handbook of Laser Welding Technologies*, pages 139–162. Elsevier, 2013. ISBN 9780857092649. doi: 10.1533/9780857098771.1.139.

Razvi, Sayyeda Saadia, Feng, Shaw, Narayanan, Anantha, Tina Lee, Yung-Tsun, and Witherell, Paul. A review of machine learning applications in additive manufacturing. 2019.

Reinke, Annika, Eisenmann, Matthias, Tizabi, Minu D., Sudre, Carole H., Rädsch, Tim, Antonelli, Michela, Arbel, Tal, Bakas, Spyridon, Cardoso, M. Jorge, Cheplygina, Veronika, Farahani, Keyvan, Glocker, Ben, Heckmann-Nötzel, Doreen, Isensee, Fabian, Jannin, Pierre, Kahn, Charles E., Kleesiek, Jens, Kurc, Tahsin, Kozubek, Michal, Landman, Bennett A., Litjens, Geert, Maier-Hein, Klaus, Menze, Bjoern, Müller, Henning, Petersen, Jens, Reyes, Mauricio, Rieke, Nicola, Stieltjes, Bram, Summers, Ronald M., Tsaftaris, Sotirios A., van Ginneken, Bram, Kopp-Schneider, Annette, Jäger, Paul, and Maier-Hein, Lena. Common limitations of image processing metrics: A picture story. 2021. URL http://arxiv.org/pdf/2104.05642v2.

Rieder, Hans, Spies, Martin, Bamberg, Joachim, and Henkel, Benjamin. On- and offline ultrasonic characterization of components built by slm additive manufacturing. AIP Conference Proceedings, page 130002. AIP Publishing LLC, 2016. doi: 10.1063/1.4940605.

Rombouts, Marleen, Kruth, J.-P., and Froyen, Ludo. Impact of physical phenomena during selective laser melting of iron powders. 2009.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. 2015. URL http://arxiv.org/pdf/1505.04597v1.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Roland J. Learning representations by back-propagating errors. *Nature*, 1986.

Saracco, Roberto. Digital transformation in manufacturing ii, 18.09.2019. URL https://cmte.ieee.org/futuredirections/2019/09/18/digital-transformation-in-manufacturing-iii/.

Schawaller, Julia. *Bildbasierte Verarbeitung von Pulverbett- und Schmelzbadaufnahmen der additiven Fertigung von Ti-6Al-4V*. Masterarbeit, Universität Hamburg, Hamburg, 2019.

Schiebold, Karlheinz. *Zerstörungsfreie Werkstoffprüfung - Ultraschallprüfung*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015a. ISBN 978-3-662-44699-7. doi: 10.1007/978-3-662-44700-0.

Schiebold, Karlheinz. *Zerstörungsfreie Werkstoffprüfung - Durchstrahlungsprüfung*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015b. ISBN 978-3-662-44668-3. doi: 10.1007/978-3-662-44669-0.

Schnars, Ulf and Kück, Andreas. Application of pod analysis at airbus. 2009.

Schörner, Karsten. *Development of Methods for Scatter Artifact Correction in Industrial X-ray Cone-beam Computed Tomography*. Dissertation, Technische Universität München, 2012.

Scime, Luke and Beuth, Jack. Using machine learning to identify in-situ melt pool signatures indicative of flaw formation in a laser powder bed fusion additive manufacturing process. 2019.

Scime, Luke, Siddel, Derek, Baird, Seth, and Paquit, Vincent. Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: A machine-agnostic algorithm for real-time pixel-wise semantic segmentation. *Additive Manufacturing*, 36:101453, 2020. ISSN 22148604. doi: 10.1016/j.addma.2020.101453.

Shevchik, S. A., Kenel, C., Leinenbach, C., and Wasmer, K. Acoustic emission for in situ quality monitoring in additive manufacturing using spectral convolutional neural networks. *Additive Manufacturing*, 21:598–604, 2018. ISSN 22148604. doi: 10.1016/j.addma.2017.11.012.

Shorten, Connor and Khoshgoftaar, Taghi M. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 2019. doi: 10.1186/s40537-019-0197-0.

Simard, Patrice Y., Steinkraus, Dave, and Platt, John C. Best practices for convolutional neural networks applied to visual document analysis. 2003.

Simonds, Brian J., Sowards, Jeffrey, Hadler, Josh, Pfeif, Erik, Wilthan, Boris, Tanner, Jack, Harris, Chandler, Williams, Paul, and Lehman, John. Time-resolved absorptance and melt pool dynamics during intense laser irradiation of a metal. *Physical review applied*, 10(4), 2018. ISSN 2331-7019. doi: 10.1103/PhysRevApplied.10.044061.

Sørensen, Thorvald. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content: Its application to analyses of the vegetation on danish commons. 1948.

Spears, Thomas G. and Gold, Scott A. In-process sensing in selective laser melting (slm) additive manufacturing. *Integrating Materials and Manufacturing Innovation*, 5(1):16–40, 2016. doi: 10.1186/s40192-016-0045-4.

Stanford University. Cs231n convolutional neural networks for visual recognition, 16.05.2022a. URL https://cs231n.github.io/convolutional-networks/.

Stanford University. Cs231n convolutional neural networks for visual recognition, 16.05.2022b. URL https://cs231n.github.io/neural-networks-1/.

Stanford University. Cs231n convolutional neural networks for visual recognition, 16.05.2022c. URL https://cs231n.github.io/neural-networks-2/.

Stock, Stuart R. *MicroComputed Tomography*. CRC Press, 2018. ISBN 9781315219189. doi: 10.1201/9781420058772.

Sudre, Carole H., Li, Wenqi, Vercauteren, Tom, Ourselin, Sébastien, and Cardoso, M. Jorge. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. 10553(11):240–248, 2017. doi: 10.1007/978-3-319-67558-9{\textunderscore}28. URL http://arxiv.org/pdf/1707.03237v3.

Thijs, Lore, Verhaeghe, Frederik, Craeghs, Tom, van Humbeeck, Jan, and Kruth, Jean-Pierre. A study of the microstructural evolution during selective laser melting of ti–6al–4v. *Acta Materialia*, 58(9):3303–3312, 2010. ISSN 13596454. doi: 10.1016/j.actamat.2010.02.004.

Vilanova, Mireia, Escribano-García, Rubén, Guraya, Teresa, and San Sebastian, Maria. Optimizing laser powder bed fusion parameters for in-738lc by response surface method. *Materials (Basel, Switzerland)*, 13(21), 2020. ISSN 1996-1944. doi: 10.3390/ma13214879.

Virkkunen, Iikka, Bolander, Martin, Myöhänen, Heikki, Myohanen, Myoehaenen, Miorelli, Roberto, Johansson, Ola, Kicherer, Philip, Curtis, Chris, and Oliver. Qualification of non-destructive testing systems that make use of machine learning (eniq recommended practice 13). 2021.

Wagner, Christian. *Untersuchungen zum selektiven Lasersintern von Metallen: Zugl.: Aachen, Techn. Hochsch., Diss., 2002*, volume 2003,11 of *Berichte aus der Produktionstechnik*. Shaker, Aachen, 2003. ISBN 3-8322-1538-7.

Wang, Zhuo, Jiang, Chen, Liu, Pengwei, Yang, Wenhua, Zhao, Ying, Horstemeyer, Mark F., Chen, Long-Qing, Hu, Zhen, and Chen, Lei. Uncertainty quantification and reduction in metal additive manufacturing. *npj Computational Materials*, 6(1), 2020. doi: 10.1038/s41524-020-00444-x.

William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, cinjon, and Anton Bakhtin. Pytorchlightning/pytorch-lightning: 0.7.6 release, 2020.

Withers, James C. and Loutfy, Raouf. Low cost processing to produce spherical titanium and titanium alloy powder, 2014.

Wong, Sebastien C., Gatt, Adam, Stamatescu, Victor, and McDonnell, Mark D. Understanding data augmentation for classification: when to warp? 2016. URL http://arxiv.org/pdf/1609.08764v2.

Wu, Bo, Ji, Xiao-yuan, Zhou, Jian-xin, Yang, Huan-qing, Peng, Dong-jian, Wang, Ze-ming, Wu, Yuan-jie, and Yin, Ya-jun. In situ monitoring methods for selective laser melting additive manufacturing process based on images — a review. *China Foundry*, 18(4):265–285, 2021. ISSN 1672-6421. doi: 10.1007/s41230-021-1111-x.

Yuan, Bodi, Guss, Gabriel M., Wilson, Aaron C., Hau-Riege, Stefan P., DePond, Phillip J., McMains, Sara, Matthews, Manyalibo J., and Giera, Brian. Machine–learning–based monitoring of laser powder bed fusion. *Advanced Materials Technologies*, 3(12):1800136, 2018. ISSN 2365-709X. doi: 10.1002/admt.201800136.

Zeng, Kai, Pal, Deepankar, and Stucker, Brent. A review of thermal analysis methods in laser sintering and selective laser melting. 2012.

Zenzinger, Guenter, Bamberg, Joachim, Ladewig, Alexander, Hess, Thomas, Henkel, Benjamin, and Satzger, Wilhelm. Process monitoring of additive manufacturing by using optical tomography. In Chimenti, Dale E. and Bond, Leonard J., editors, *41ST ANNUAL REVIEW OF PROGRESS IN QUANTITATIVE NONDESTRUCTIVE EVALUATION: Volume 34*, AIP Conference Proceedings, pages 164–170. AIP Publishing LLC, 2015. doi: 10.1063/1.4914606.

Zhang, Bi, Li, Yongtao, and Bai, Qian. Defect formation mechanisms in selective laser melting: A review. *Chinese Journal of Mechanical Engineering*, 30(3):515–527, 2017. ISSN 1000-9345. doi: 10.1007/s10033-017-0121-5.

Zhang, Daniel, Mishra, Saurabh, Brynjolfsson, Erik, Etchemendy, John, Ganguli, Deep, Grosz, Barbara, Lyons, Terah, Manyika, James, Niebles, Juan Carlos, Sellitto, Michael, Shoham, Yoav, Clark, Jack, and Perrault, Raymond. The ai index 2021 annual report. 2021.

Publications and Conferences

Publications

Jonas Holtmann, Denis Kiefel, Stefan Neumann, Rainer Stoessel, and Christian U. Grosse. A data driven approach to the online monitoring of the additive manufacturing process. *Advanced Materials Research*, 1161:137-144, 2021a. doi: 10.4028/www.scientific.net/ AMR.1161.137 (peer-reviewed)

Jonas Holtmann, Jordy Permadi, Denis Kiefel, and Christian Grosse. Convolutional neural networks for the automated segmentation of computed tomography scans. *Special Issue of e-Journal of Non-destructive Testing (eJNDT)*, (VOL.27(6)), 2021b.

Björn Milcke, Pascal Dinglinger and Jonas Holtmann. Exploring the Role of Explainable AI in the Development and Qualification of Aircraft Quality Assurance Processes: A Case Study. *Explainable Artificial Intelligence*, (Volume 2156 of Communications in Computer and Information Science), Pages 331–352. Springer Nature Switzerland, 2024. ISBN 978-3-031-63802-2 . doi: 10.1007/978-3-031-63803-9 _18 . (peer-reviewed)

Conferences

Jonas Holtmann. A data driven approach to the online monitoring of the additive manufacturing process. International Symposium. Material Science and Technology of Additive Manufacturing. Bremen. 2019.

Jonas Holtmann. Automated process monitoring of the metal additive manufacturing process – time and cost savings thanks to neural networks. Aerospace Technology Week. Conference on Testing. 2021. Toulouse.

Jonas Holtmann. Convolutional neural networks for the automated segmentation of computed tomography scans. The 13th International Symposium on NDT in Aerospace. 2021. Williamsburg, VA.

Supervised and Supported Student Theses

Julia Schawaller. Bildbasierte Verarbeitung von Pulverbett- und Schmelzbadaufnahmen der additiven Fertigung von Ti-6Al-4V. Master Thesis. 2019.

Paulo Montijo Bandeira. Data Registration of Computed Tomography and Online Monitoring Data in Additive Manufacturing. Master Thesis. 2020.

Sarvesh Satish Shimpi. Comparison of Image Analysis Algorithms for Computed Tomography Images of Additive Manufactured Components. Master Thesis. 2020.

Raffael Kaehn. Digital Image Processing for In-Situ Defect Detection in Laser Powder Bed Fusion. Bachelor Thesis. 2021.

Jordy Luberizky Permadi. Automated Anomaly Detection in Computed Tomography Scan Images of Additive Manufactured Titanium Components using Deep Learning. Master Thesis. 2021

Martin Lotz. Analyse von Online-Monitoring-Daten aus der additiven Fertigung mittels neuronaler Netze. Master Thesis. 2022

Imko Schumacher. Evaluation and Implementation of Rigid and Deformable 3D Registration in Additive Manufacturing. Bachelor Thesis. 2022.

Björn Milcke. Explainability of Machine Learning Algorithms in Quality Assurance Based on the Example of Inline Monitoring for Additive Manufacturing. Master Thesis. 2022.

List of Abbreviations

AI	Artificial Intelligence
AM	Additive Manufacturing
BFN	Biggest False Negative
BFP	Biggest False Positive
BUD	Biggest Undetected Defect
CC	Connected Component
CNN	Convolutional Neural Network
CT	Computed Tomography
EASA	European Union Aviation Safety Agency
FN	False Negative
FNN	Feedforward Neural Network
FP	False Positive
FP_{400}	Number of False Positives larger than 400 µm
GDL	Generalized Dice Loss
GT	Ground Truth
IR	Infrared
LPBF	Laser Powder Bed Fusion
LSS	Label Shape Statistics
ML	Machine Learning
NDT	Non-Destructive Testing
nifti	Neuroimaging Informatics Technology Initiative
NUD	Number of Undetected Defects
NUD_{400}	Number/ Percentage of Undetected Defects larger than 400 µm
OM	Online Monitoring
OS	Optical System
OT	Optical Tomography
PBF-LB/M	Powder Bed Fusion - Laser Based/ Metal

POD . . . . . Probability of Detection

SGLB . . . . . Sigma Labs

SVM . . . . . Support Vector Machine

TED . . . . . Thermal Emission Density

TEP . . . . . Thermal Emission Planck

UT . . . . . Ultrasonic Testing

xAI . . . . . Explainable Artificial Intelligence

Appendix

1 Buildjob E

The Buildjob E contains more complex samples. It consists of five propellers[1], two pyramids[2] and a model of the *Elbphilharmonie*[3] in Hamburg. Two propellers (Propeller 2 and 3) were printed with standard parameters, two (Propeller 4 and 5) with reduced focus (approximately 40%) and one propeller (Propeller 1) with the lack of fusion parameter set. No inclusions were induced on purpose by powder contamination. Pyramid 1 was also produced with lack of fusion parameters while Pyramid 2 was printed with standard parameters. The *Elphi* was printed with standard parameters but was put together from multiple CAD files which lead to imperfect slicing and sub-optimal hatching of the specimen.

Figure 1 Specimens of the Buildjob E with five propellers, two pyramids and a model of the *Elbphilharmonie*.

2 Buildjobs 100 - 700

The Buildjobs 100 to 500 were printed before the Buildjobs A to E. They consisted of the same geometry and similar parameter sets and were intended for the same use. The data gathered during these prints could not be used for the study as the installed sensor system had a defect. The sensor supplier was not able to correct the recorded data. Therefore, the buildjobs had to be printed again. However, as the produced specimens were CT scanned the CT data obtained for the samples could be used for the study. In particular, the data could be used for the training the CT CNN.

[1] adapted from Propeller by drayde on https://www.thingiverse.com/thing:3678
[2] adapted from Pyramid of Kukulcan at Chichen Itza by gpvillamil on https://www.thingiverse.com/thing:4155
[3] adapted from Elbphilharmonie by BjBerry00 on https://www.thingiverse.com/thing:4600376/files

Buildjob 100

The entire buildjob was printed with a reduced gasflow. This may lead to more soot which may influence the laser beam and in extreme cases stain the laser optics. Overall this may lead to a more unstable process with more anomalies. The gasflow was decreased gradually from the standard flow down to the allowed minimum flow of the printer. The increase in soot during the printing was visible with the naked eye. Overall a total of 20 specimens was printed. However, the post-process CT inspection did not reveal any relevant anomalies in the specimens.

Buildjob 200

The Buildjob 200 consists of different parameter sets. The first 18 specimens were printed with an increased focus where the focus was increased incrementally every three samples. The first three samples were printed with the standard focus where as the last three samples were printed with an increase in focus of around 60 %. For the highest focus increase the CT inspection showed a large number of very small pores most being below the relevant detection size of $400\,\mu m$. The last twelve specimens on the buildplate were printed with standard parameters and did not show any relevant anomalies. Therefore, the parameter set was deemed unsuitable for the creation of relevant reference anomalies.

Buildjob 300

For Buildjob 300 the dosing factor per layer was decreased in combination with a low gas flow. The dosing factor describes the amount of powder used per layer. A dosing factor of 100 % indicates that the amount of powder dispersed by the recoater is exactly equal to the amount needed mathematically to fill the buildplate by the specified layer thickness. As there exist a variety of imperfections during the recoating process this volume is normally multiplied by a safety factor. E.g. this takes into account that some powder might be lost during the recoating. As the recoater disperses the powder from right to left on the buildplate a possible shortage of powder impacts the left of the buildplate more strongly. In the case of missing powder the laser cannot melt the powder to create a new layer but instead remelts the layer below. As the dosing factor was not decreased below 100 % the existence of powderless regions was not guaranteed but the statistical likelihood was increased and hence the creation of anomalies.

The effect of the powder shortage was clearly visible in the CT inspection. In the proximity of layers with a low dosing factor large porous regions could be observed. Those were limited to a few layers in height but did expand significantly in the xy-plane. As expected the effect was more pronounced for the samples on the left of the buildplate.

Buildjob 400

Buildjob 400 was not printed.

Buildjob 500

Buildjob 500 is identical to the Buildjob B described in Section 4.1 with the exception of the stepsize by which the laser focus or the skywriting delay was decreased. For Buildjob 500 the focus was decreased gradually for every three samples in steps between approximately 7 % and 15 %. The skywriting delay was decreased in steps of 10 %. The CT scan showed anomalies in most of the printed specimens. Furthermore, an increase in relevant anomalies for the more severe deviations in focus and skywriting delay could be observed. The anomalies could be well distinguished and did not hinder the printing of the part. As a lesson learnt the Buildjob B focused more on the strongly altered parameter sets with a higher likelihood for anomaly creation as those samples contain a higher number of anomalies while still allowing for a robust building process.

Buildjob 600

Buildjob 600 combines specimens printed with the previously introduced parameter set in Section 4.1 and specimens printed with the standard parameter.

Buildjob 700

Buildjob 700 is a duplicate of Buildjob 500.

Buildjob X

In contrast to the previously described buildjobs the Buildjob X does not aim to produce pores as anomalies but instead focuses on the creation of inclusions in the specimens (similar to Section 4.1). For this purpose tungsten powder was added to the titanium powder in the powder reservoir so that tungsten powder makes up around 2 % of the total powder weight . In contrast to Buildjob C the powder was subsequently mixed by hand resulting in a more uniform distribution of the tungsten particles in the reservoir and hence over the buildheight.