Deep Learning-Based Anomaly Detection for MR-Only Radiotherapy

Towards Automatic Synthetic CT Quality Assurance for MR-Only Radiotherapy

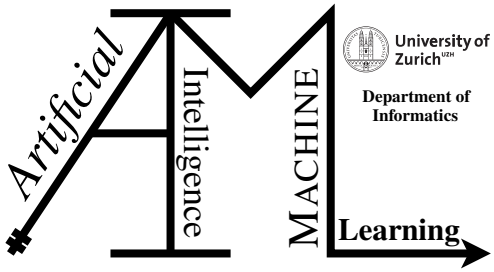
Master Project

Ömer Yildirim, Justin Verhoek, Weijie Chen, Yeyang Liu, Kai Koepchen

24-746-372, 18-750-661, 23-743-727, 24-745-374, 24-738-189

Submitted on
December 25, 2025

Thesis Supervisor
Prof. Dr. Manuel Günther
Mariia Lapaeva, Dr. Riccardo Dal Bello, Dr. Stephanie Tanadini-Lang



Declaration of Independence for Written Work

I hereby declare that I have **composed** this work independently and without the use of any aids other than those declared (including generative AI such as ChatGPT) – the use of generative AI to **improve** my composed work was permitted by the thesis supervisor. I am aware that I take full responsibility for the scientific character of the submitted text myself, even if AI aids were used. All passages taken verbatim or in sense from published or unpublished writings are identified as such. The work has not yet been submitted in the same or similar form or in excerpts as part of another examination.

Zurich, Switzerland, 24.12.2025	
Place, Date	Ömer Yildirim
Zurich, Switzerland, 24.12.2025	
Place, Date	Justin Verhoek
Zurich, Switzerland, 24.12.2025	
Place, Date	Weijie Chen
Zurich, Switzerland, 23.12.2025	liu Yeyang.
Place, Date	Yeyang Liu
Zurich, Switzerland, 24.12.2025	
Place, Date	Kai Koepchen

Master Project

Author: Ömer Yildirim, Justin Verhoek, Weijie Chen, Yeyang Liu, Kai Koepchen

Project period: 19.06.2025 - 25.12.2025

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

We sincerely thank all who contributed to this project. Our special thanks go to Mariia Lapaeva for her constant support, weekly discussion, and for introducing us to essential medical imaging knowledge. Her dedication and insight were key to completing this research. We also express our sincere gratitude to Professor Dr. Manuel Günther, for his guidance on the detailed feedback. We are also grateful for the computing resources provided by the AIML group, which made our experiments possible. We extend our thanks to Dr. Riccardo Dal Bello and Dr. Stephanie Tanadini-Lang for their supervision and valuable insights throughout this research. Finally, we thank our teammates for their collaboration and shared efforts, which made this research both productive and rewarding.

Abstract

This report investigates deep learning-based anomaly detection as a patient-specific quality assurance step for MR-only radiotherapy, focusing on detecting metal-induced artifacts in pelvic MR images before synthetic CT generation. The study evaluates multiple unsupervised and one-class approaches, including reconstruction-based, knowledge distillation-based, flow-based, and memory-bank methods, on a curated subset of the SynthRAD2023 pelvis dataset. The models are compared across pixel-, slice-, and patient-level metrics, and the impact of input representations (replicated MR channels, 2.5D adjacent slices, and bone colormaps) as well as ImageNet versus RadImageNet pretraining for FastFlow is analyzed. Among all evaluated configurations, the memory-bank method PatchCore achieves the most favorable trade-off between localization and patient-level detection, reaching a post-processed best patient-level Dice of 0.90 on the single-center setting, and remaining comparatively robust under multi-center domain shifts. These findings suggest that anomaly detection on MR alone can support automatic pre-screening for MR-only workflows and that PatchCore-style memory methods are a promising component for future clinically oriented MR-only QA pipelines, although further prospective and multi-center validation remains necessary.

Contents

Chapter 1

Introduction

One of the recent advances in the field of radiation oncology is Magnetic Resonance (MR)-only Radiotherapy (RT), a workflow evolving from the current practice of MR-guided Radiation Therapy (MRgRT) (Chin et al., 2020; Hall et al., 2019). In a conventional MRgRT workflow, planning involves multiple imaging modalities. MR imaging is leveraged for its superior soft-tissue contrast, which is highly advantageous for accurately delineating the tumor target and surrounding organs-at-risk (OARs) (Chin et al., 2020; Hall et al., 2019). On the other hand, a Computed Tomography (CT) scan is acquired because it provides the electron density (ED) information, in Hounsfield Units (HU), that is essential for accurate radiation dose calculation (Spadea et al., 2021). This conventional approach, however, requires a complex registration step to fuse the MR (used for delineation) with the CT (used for planning), a process that can introduce geometric uncertainties (Spadea et al., 2021). Furthermore, the CT scan exposes the patient to additional ionizing radiation (Spadea et al., 2021).

The MR-only workflow aims to overcome these challenges by eliminating the need for a separate CT scan (Spadea et al., 2021; Lapaeva et al., 2022; Dal Bello et al., 2023). The benefits of this approach are significant: it reduces the patient's total radiation exposure, simplifies the clinical workflow, potentially reduces healthcare costs, and eliminates the geometric uncertainties associated with CT-MR registration (Spadea et al., 2021; Placidi et al., 2020). This workflow is enabled by generating a synthetic CT (sCT) directly from the MR data, typically using deep learning methods such as neural networks (Spadea et al., 2021; Lapaeva et al., 2022; La Greca Saint-Esteven et al., 2023). Modern architectures, such as Generative Adversarial Networks (GANs) and Vision Transformers (ViTs), approach this as an image-to-image translation task, learning to map MR signal intensities to the required electron density values (Spadea et al., 2021; Lapaeva et al., 2022). These data-driven approaches generally outperform classical atlas-based methods, producing sCTs with high structural fidelity and dosimetric accuracy (Spadea et al., 2021).

The quality and reliability of the sCT generation process are critical for the clinical feasibility of MR-only RT (Dal Bello et al., 2023; Nella et al., 2025; van Harten et al., 2020). While much research in the field of patient-specific quality assurance (PSQA) focuses on retrospectively evaluating the accuracy of a generated sCT, such as comparing its radiation dose distribution against a real CT scan, this approach has limitations (Dal Bello et al., 2023; Nella et al., 2025; Chourak et al., 2022a). Such studies are often performed retrospectively, are specific to a single treatment center, can involve time-consuming manual manipulations, and, most importantly, check for errors *after* the sCT has already been created (van Harten et al., 2020; Dal Bello et al., 2023; Nella et al., 2025).

A major challenge remains: performing automated QA *before* the sCT generation (van Harten et al., 2020; Parchur et al., 2025). This is particularly crucial when the input MR images contain anomalies. In this work, the most critical anomalies are metal implants, which induce severe susceptibility artifacts in MR images. In this work, the most critical anomalies are metal implants, which induce severe susceptibility artifacts in MR. These artifacts manifest as signal

voids, pile-ups, and geometric distortions that typically extend well beyond the physical extent of the implant, resulting in artifact regions larger than the underlying metal objects (Hargreaves et al., 2011; Chen et al., 2011). Most sCT generation studies are conducted in small, curated patient cohorts that explicitly exclude cases with metal implants, significant imaging artifacts, or other unusual anatomical patterns (Lapaeva et al., 2022; La Greca Saint-Esteven et al., 2023). Consequently, when a neural network encounters such out-of-distribution (OOD) inputs, which it has not seen during training, the quality of the generated sCT can be severely compromised (Parchur et al., 2025; Law et al., 2024). This can lead to inaccurate dose calculations and potentially unsafe treatment plans (Dal Bello et al., 2023; Li et al., 2024; Hémon et al., 2025). Therefore, a vital step toward making MR-only RT clinically viable is developing an automatic quality assurance pipeline that can identify and exclude out-of-distribution MR images *before* the sCT generation stage (van Harten et al., 2020; Parchur et al., 2025), thereby improving cost-effectiveness and workflow efficiency for MR-only RT.

The primary aim of this research is to investigate the feasibility of using automated anomaly detection (AD) methods to support PSQA for MR-only radiotherapy. This involves detecting anomalies in MR images prior to the sCT generation stage. In the best-case scenario, such a system could highlight anomalies while the patient is still undergoing the MR simulation (van Harten et al., 2020). Patients identified with unsuitable MR images could then be directed to undergo conventional CT scanning to ensure safe and effective treatment planning (Nella et al., 2025). Conversely, patients whose images are confirmed to be free of OOD features could be confidently directed to the MR-only workflow (Nella et al., 2025). While anomaly and OOD detection models have been successfully applied in other fields, their application as a pre-screening tool in the MR-only sCT pipeline is not well-established (Bao et al., 2024; Akcay et al., 2022; Parchur et al., 2025). Therefore, the goal of this study is to explore and evaluate different AD methodologies for this specific task, laying the groundwork for a more robust and automated QA pipeline.

This master's project report is organized as follows: Chapter 2 provides the necessary **Clinical Background** on MR and CT imaging, as well as the MR-guided and MR-only radiotherapy workflows. Chapter 3 reviews **Related Work** in PSQA and OOD detection for medical imaging. Chapter 4 details the **Technical Background** of the anomaly detection approaches explored in this work. Chapter 5 describes the **Materials and Methods**, including the datasets, annotation strategy, and the post-processing pipeline. Chapter 6 outlines the **Experiments** conducted to evaluate the different AD models. The **Results** are presented in Chapter 7 and subsequently analyzed in the **Discussion** in Chapter 8 with suggestions for future work. Finally, Chapter 9 offers a **Conclusion** of the research findings.

Chapter 2

Clinical Background

2.1 Magnetic Resonance (MR) Imaging

MR imaging is a noninvasive imaging modality that generates high-resolution, volumetric images of the body by utilizing the magnetic properties of hydrogen protons (Weishaupt et al., 2007). In the context of radiation oncology, its primary advantage is the superior soft-tissue contrast it provides (Chandarana et al., 2018). This property is highly beneficial for accurately delineating the tumor target and surrounding organs-at-risk (OARs), which is a key step in treatment planning (Jonsson et al., 2019).

While MR images offer high-quality anatomical detail, they do not directly provide the electron density (ED) information used for radiation dose calculation (Wang et al., 2008). Furthermore, unlike CT, MR signal intensities are not absolute measurements but are influenced by scanner-specific characteristics and acquisition parameters. Consequently, preprocessing strategies such as intensity normalization are required to standardize the data for automated processing.

For the purposes of this master's project, the relevant aspect of MR is how different materials and tissues are represented. While soft tissues are well-differentiated, other structures, such as bone or air, typically appear as dark regions with low signal. In some cases, contrast agents such as gadolinium are administered to enhance the visibility of specific tissues, appearing as a distinct brightening of vascular structures. Metallic implants, a key out-of-distribution (OOD) class for this study, also produce distinct visual patterns of characteristic artifacts, such as dark signal voids (regions of no signal) and local geometric distortions (Hargreaves et al., 2011). These MR-visible artifacts and enhancement patterns are the signals that the anomaly detection models in this work are trained to identify, as they represent patterns outside of the in-distribution anatomical data used for sCT generation.

2.2 Computed Tomography (CT) Imaging

Computed Tomography (CT) is a radiological modality in modern clinical practice, providing high-resolution, cross-sectional views of internal anatomy (Jung, 2021; Goldman, 2007). In contrast to conventional X-ray radiography, which generates two-dimensional superimposed images, CT enables three-dimensional visualization of tissues, facilitating comprehensive assessment of pathology and guiding therapeutic decision-making (Jung, 2021). CT image formation relies on measuring the attenuation of X-ray beams as they traverse tissues with differing physical densities. The attenuation coefficient, which quantifies the reduction in X-ray intensity, is influenced

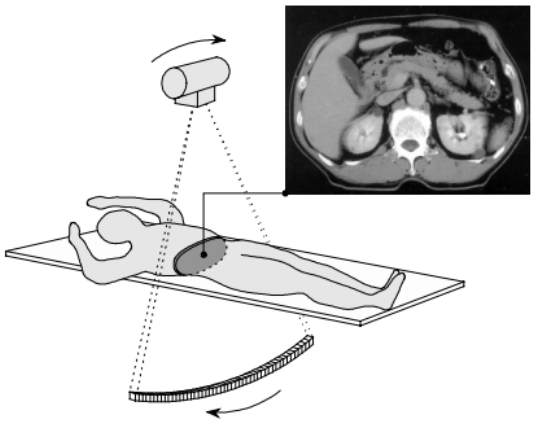


Figure 2.1: SCHEMATIC ILLUSTRATION OF A CT IMAGING PROCESS. *The gantry rotates around the patient lying on a motorized table, acquiring projection data used to generate cross-sectional images of internal anatomy (Michael, 2001).*

primarily by the photoelectric effect and Compton scattering—both dependent on electron density (Jung, 2021; Goldman, 2007). CT scanners acquire transmission data from multiple angles by rotating the gantry, which contains both the X-ray tube and detectors, around a patient positioned on a motorized table. The process is depicted schematically in Figure 2.1, showing the principles of data acquisition and tomographic slice reconstruction. The acquired signals are then reconstructed into high-resolution images using specialized algorithms (Jung, 2021; Goldman, 2007). All CT images are standardized in Hounsfield Units (HU), a scale reflecting the degree of X-ray attenuation for each tissue type. Water is defined as 0 HU, air as -1000 HU, and bone can range from +1000 to +2000 HU. This standardized quantification allows reliable differentiation and comparison of anatomical structures in clinical workflows (Goldman, 2007). CT's speed and accessibility, coupled with its robust anatomical coverage, make it indispensable in diagnostic and therapeutic planning (Pereira et al., 2014).

In the context of radiotherapy, CT imaging offers the advantage of providing direct electron density information mapped from HU values. These quantitative measurements are critical for precise calculation and delivery of radiotherapy dose to the tumor, minimizing exposure to surrounding healthy tissues (Pereira et al., 2014).

2.3 MR-Guided vs. MR-Only Radiotherapy

Comparison of MR-Guided and MR-Only Workflows

MR-Guided Radiotherapy (MRgRT) integrates MR imaging with conventional CT-based treatment planning. Modern MR-Linac systems combine the two modalities, enabling adaptive treatment adjustments based on anatomical changes observed during therapy. In MRgRT, the planning CT and MR must be co-registered so that MR-based contours align with the CT dose grid (Nyholm et al., 2009). This registration typically matches anatomical landmarks or implanted markers. However, even with advanced algorithms, CT–MR registration can introduce spatial uncertainty (Nyholm et al., 2009).

MR-Only Radiotherapy, in contrast, relies exclusively on MR imaging for both treatment planning and guidance. An sCT is generated from MR data to estimate the electron density for

dose calculation (Johnstone et al., 2018), which can be challenging for low-signal regions such as bone or air (Paradis et al., 2015). Accurate bone-air differentiation is essential to maintain dosimetric precision. Therefore, assessing the integrity of input MR images before sCT generation becomes an important quality assurance (QA) step.

Challenges of Transitioning to an MR-only Workflow

MR-only radiotherapy eliminates CT-MR registration uncertainties, reduces ionizing radiation exposure, and streamlines clinical workflows. However, as described in the Chapter 1, transitioning to an MR-only workflow introduces technical considerations related to MR-specific distortions and artifacts that must be managed.

MR imaging is inherently sensitive to magnetic field inhomogeneities (Walker et al., 2022), gradient nonlinearities (Höfler et al., 2025), and susceptibility effects (Wang et al., 2013), especially at air–tissue interfaces or near metallic implants. These effects can introduce minor geometric distortions or local signal voids. Metal implants cause particularly severe artifacts in MR, appearing as signal voids or pile-ups that are often indistinguishable from air bubbles or bones, resulting in an artifact region that is typically larger than the actual metal object. Although clinical calibration minimizes such errors, residual distortions can influence anatomical accuracy in specific cases.

Most current QA procedures still depend on reference CT or sCT verification to confirm dosimetric accuracy (Dal Bello et al., 2023). Routine QA includes geometric distortion checks using phantoms and voxel-wise dose comparisons between sCT and CT-based plans (Chourak et al., 2022b). Recently, unsupervised deep-learning methods have been proposed to automate early MR image screening (Kim et al., 2021; Vasiliuk et al., 2023). These algorithms flag OOD MR scans that contain atypical artifacts or anatomy before sCT generation. Because they do not require labeled examples of every anomaly, such methods can efficiently identify unusual scans (e.g., unexpected implants or severe artifact) as an initial QA filter. Integrating such automated OOD checks can enhance workflow reliability while maintaining clinical efficiency.

Chapter 3

Related Work

The clinical implementation of an MR-only RT workflow relies on generating an sCT from MR data to derive the electron density information required for dose calculation (Lapaeva et al., 2022; Spadea et al., 2021). To facilitate this process, modern DL architectures, from Generative Adversarial Networks (GANs) to Vision Transformers (ViTs), have demonstrated capabilities in generating sCTs. Several studies indicate that these methods can achieve high dosimetric accuracy, which assesses the agreement of dose distributions re-calculated on the sCT against reference CT plans, often with dose deviations below the 2% clinical threshold (La Greca Saint-Esteven et al., 2023) and frequently under 1% (Spadea et al., 2021; La Greca Saint-Esteven et al., 2023). To benchmark these capabilities, the SynthRAD2023 challenge evaluated sCT generation methods on a large multi-center dataset, reporting that top-performing algorithms achieved high structural similarity and gamma pass rates for photon and proton plans (Huijben et al., 2024).

sCT generation approaches have also been tailored for MR-Linac adaptive treatment delivery workflows (Ocanto et al., 2024). In this context, Lapaeva et al. (2022) successfully used a CycleGAN to handle the abdominal region on a low-field MR-Linac, correctly mapping mobile air pockets from unpaired MR images and achieving mean dose-volume histogram (DVH) deviations below 1%. Similarly, La Greca Saint-Esteven et al. (2023) applied a Residual Vision Transformer (ResViT) model for low-field sCT generation in the head-and-neck (HN) region, also reporting DVH deviations below 1%.

Despite these advancements, majority of sCT research and clinical studies exclude patients with imaging artifacts, unseen anatomy or implants, which could affect neural network performance and the quality of the sCT, making them unsuitable for treatment planning (Fusella et al., 2024). Consequently, the application of neural networks to out-of-distribution inputs may yield sCTs with diminished dosimetric or geometric accuracy relative to in-distribution performance.

The clinical adoption of MR-only radiotherapy depends on the accuracy of the sCT images used for dose calculation (Law et al., 2024; Li et al., 2024). Discrepancies in the sCT, such as tissue misclassification or hallucinations, require careful management to maintain dosimetric accuracy (Parchur et al., 2025; Zaffino et al., 2024). Current PSQA approaches are often retrospective and either require real CTs to measure image similarity and dosimetric differences or estimate uncertainty after the sCT generation step (Zaffino et al., 2024; Chourak et al., 2022a; van Harten et al., 2020). While these methods are effective, they can involve manual effort and limit real-time application in scenarios where the ground-truth (GT) planning CT is unavailable.

A primary line of PSQA research focuses on verifying the dosimetric accuracy of a generated sCT using independent models. This approach validates the final sCT-based dose plan by recalculating it on an alternative, trusted electron density (ED) map. Nella et al. (2025) established a practical, tiered clinical workflow for this purpose. Their PSQA protocol begins with a swift and simple check (Method A), recalculating the dose on a uniform water-override body mask, which assigns a constant water density to the entire anatomical volume to verify consistency indepen-

dent of tissue heterogeneities. If this fails a 5% deviation criterion, a more complex (Method B) check using bulk-density overrides for different tissue classes (e.g., bone, fat, air) is performed. If both (A) and (B) fail, the workflow mandates a fall-back to a conventional CT scan (Method C) (Nella et al., 2025; Dal Bello et al., 2023). Dal Bello et al. (2023) also investigated this comparative strategy, assessing water, bulk densities, and a deformed planning CT (dCT). As illustrated in Figure 3.1, the inclusion of the dCT in the comparison allows for the identification of OOD features, such as artificial implants, which may remain invisible on the MR simulation and the derived sCT. Notably, they introduced the concept of using an sCT generated by a separate, independent neural network as a PSQA tool. Their findings suggest that comparing the dose calculations from two independent sCT generators is a viable and robust strategy that can meet the 2% clinical tolerance, eliminating the need for manual bulk density contouring (Dal Bello et al., 2023).

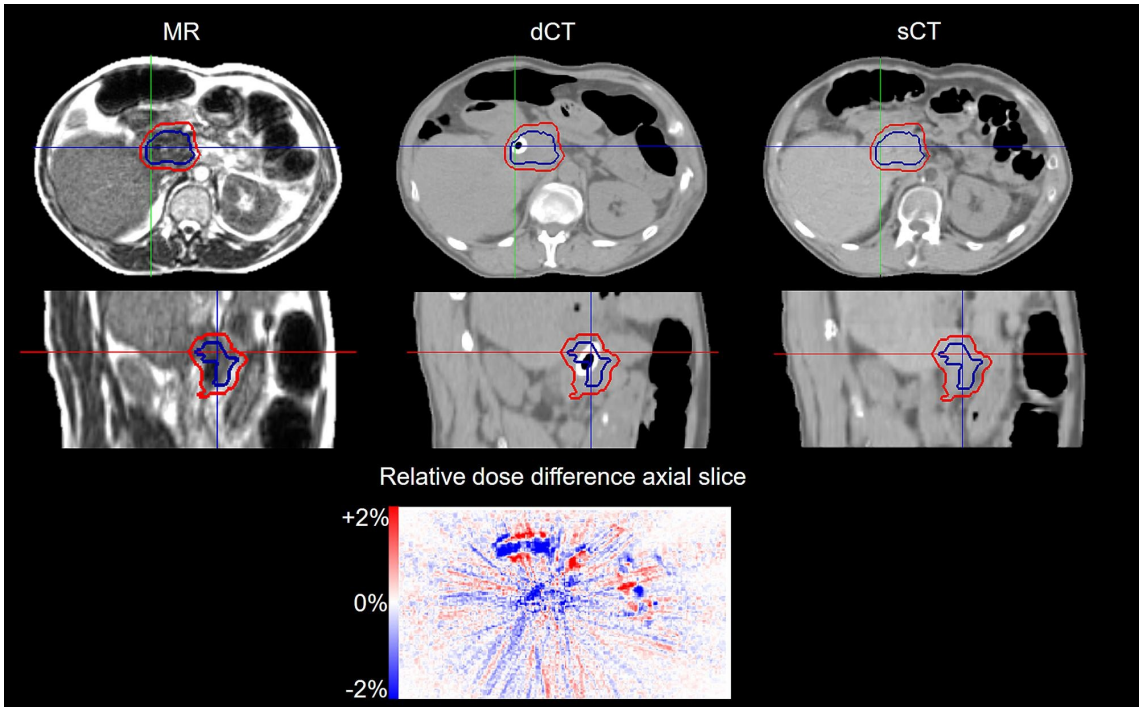


Figure 3.1: EXEMPLARY CASE OF SCT GENERATION IN THE PRESENCE OF AN ARTIFICIAL IMPLANT. *The implant is not visible on the MR simulation (left) and not reproduced in the sCT (right), representing an out-of-distribution case. The dCT (centre) identifies the voxels with higher density. The PTV (red) and GTV (blue) contours are reported alongside the relative dose difference between the reference sCT and the dCT for the axial slice (Dal Bello et al., 2023).*

Another interesting approach for PSQA methods focuses on identifying specific, localized errors within the sCT. Parchur et al. (2025) developed an "automated hallucination screener" specifically to detect false bone structures in pelvic sCTs. Their method trains a deep learning autosegmentation (DLAS) model to identify bone directly from the input MR image. This MR-based bone segmentation is then compared to the bone regions present on the generated sCT, which are identified via HU thresholding, leveraging the high attenuation values characteristic of bone structures to distinguish them from surrounding soft tissue. Any spatial discrepancy between the two-i.e., bone on the sCT that has no corresponding structure on the MR-is flagged as a halluci-

nation (Parchur et al., 2025). Similarly, Chourak et al. (2022a) proposed a voxel-wise local quality assessment by creating a reference atlas from a cohort of patient CTs. This atlas provides a mean CT image and a standard deviation map for the population. A new sCT is then assessed by registering it to the atlas and computing a z-score for each voxel, highlighting local outliers that deviate from the learned anatomical norm (Chourak et al., 2022a). Zaffino et al. (2024) extended this concept by training a DL model to predict the Mean Absolute Error (MAE) of an sCT slice. In this framework, the generated sCT slice is fed into a cascade of neural networks (a classifier for a raw MAE interval, followed by a class-specific regressor) that outputs a final, predicted MAE value (Zaffino et al., 2024). One limitation of these approaches might be that the golden truth for the evaluation of synthetic CT quality remains dosimetry metrics. As found by the SynthRAD2023 challenge (Fusella et al., 2024) and Lapaeva et al. (2025), there is no strong correlation in the area of high dosimetric differences between DVH metrics and image similarity metrics. Moreover, those approaches rely on assessments of the output sCT, rather than the input MR directly.

Another strategy for PSQA is to integrate quality assurance directly into the sCT generation process through uncertainty quantification (UQ). Rather than a separate verification step, the generation model itself estimates its own confidence. Several studies have implemented this using Bayesian Neural Networks (BNNs). Li et al. (2024) developed an uncertainty-aware framework where a BNN predicts both the sCT and a corresponding voxel-wise uncertainty map, which was found to positively correlate with the absolute prediction error. This uncertainty information was then directly incorporated into the robust optimization of proton therapy plans (Li et al., 2024). Likewise, Law et al. (2024) used a Bayesian deep network to simultaneously generate the sCT and its uncertainty. They established a linear correlation between the Bayesian model uncertainty and the actual sCT error (MAE) (Law et al., 2024). However, as noted regarding image similarity metrics, the correlation with dosimetric accuracy remains the primary verification standard.

An alternative to Bayesian UQ is the use of deep ensembles, as demonstrated by van Harten et al. (2020). They trained an ensemble of three separate sCT generators: on axial, coronal, and sagittal 2D slices, respectively. The final sCT is the voxel-wise median of the three outputs, and the uncertainty map is derived from their disagreement (e.g., the maximum voxel-wise difference) (van Harten et al., 2020). While voxel-wise median aggregation improves robustness against outliers, it may inadvertently introduce a degree of image blurring, potentially smoothing over fine anatomical details present in the individual predictions. This ensemble disagreement was reported to correlate with image similarity metrics (van Harten et al., 2020). Furthermore, this method demonstrated effectiveness for detecting OOD input images. The ensemble uncertainty was higher for MR scans acquired with a different scanner (OASIS set) or with gadolinium contrast compared to the in-distribution training data (van Harten et al., 2020).

While methods exist for post-generation dosimetric checks (Nella et al., 2025; Dal Bello et al., 2023), error detection (Parchur et al., 2025; Chourak et al., 2022a), and concurrent uncertainty estimation (Law et al., 2024; Li et al., 2024; van Harten et al., 2020), these approaches generally operate after the sCT has been generated and typically rely on the availability of a CT or the generated sCT itself. Even the ensemble method, which can detect OOD inputs, derives this information from the processing of multiple generative networks (van Harten et al., 2020). In contrast, Lapaeva (2025) introduced a proof-of-concept for anomaly detection on the step before sCT generation, based on MR images only. Their findings suggest that deep learning models have the potential to support real-time MR-only PSQA by highlighting spatial anomalies without relying on any CT or sCT, thereby facilitating the screening of out-of-distribution cases directly on input MR images before the sCT generation stage (Lapaeva, 2025).

Chapter 4

Technical Background

Anomaly detection (AD) aims to identify image patterns that deviate from the *in-distribution* patterns learned by a model during training. In medical imaging, this typically means highlighting voxels or slices that do not conform to the anatomy, appearance, or acquisition statistics learned from the in-distribution training data. In the context of this master's project, **in-distribution data** refers to MR images deemed suitable for sCT generation, covering standard anatomy as well as typical pathologies such as tumors. **Out-of-distribution data** encompasses cases unsuitable for this workflow, specifically images containing metallic implants, imaging artifacts, or atypical anatomical patterns, such as missing bone structures resulting from surgical resections. A practical reason for using AD in MR-only radiotherapy is to identify such atypical MR images earlier in the clinical workflow. Because out-of-distribution cases are highly diverse, it is often not feasible to collect an exhaustive set of labeled examples covering all possible deviations. Unsupervised settings, where models learn only from in-distribution data, are therefore commonly chosen.

In this master's project, we consider several distinct categories of AD approaches as outlined by Bao et al. (2024). **Reconstruction-Based Anomaly Detection** involves training a model on in-distribution data (MR images without anomalies) so that it produces a reconstruction representative of this data; OOD patterns are indicated by discrepancies between the input and its reconstruction. This includes autoencoders, masking/inpainting, and more recent transformer/diffusion reconstructors (Zhou and Paffenroth, 2017; Zavrtanik et al., 2020; You et al., 2022b; Teng et al., 2022). Another approach is **One-Class Classification**, which learns a compact region for in-distribution data in feature space and flags samples outside it. Typical examples are hypersphere or hyperplane formulations such as DeepSVDD (Ruff et al., 2018).

Alternative strategies focus on feature consistency or density estimation. **Knowledge Distillation Models** use a teacher-student setup trained on in-distribution data; the student should match the teacher on in-distribution inputs but deviates on out-of-distribution ones. The deviation becomes the anomaly signal, with RD4AD being a common representative (Deng and Li, 2022). Furthermore, **Memory-Bank Methods** store a set of in-distribution feature prototypes and detect anomalies by calculating the distance to the nearest neighbor within this memory. PatchCore is a widely used approach in this family (Roth et al., 2022). Finally, **Flow-Based Methods** model the likelihood of in-distribution features via normalizing flows and treat low-likelihood regions as anomalous. CFlow is a representative example (Gudovskiy et al., 2022).

4.1 Reconstruction-Based Anomaly Detection

Reconstruction-based AD assumes a model can learn how an in-distribution MR image should look and then fail on parts that are out-of-distribution. Let $\mathbf{I}$ be a 2D input slice and g_θ a re-

constructor trained only on in-distribution data. The output $\hat{\mathbf{I}} = g_\theta(\mathbf{I})$ is expected to be an in-distribution-looking version of $\mathbf{I}$. Where the input departs from in-distribution anatomy or signal statistics, the reconstruction usually cannot match it well. That gap is what we use as evidence for an anomaly.

Most methods in reconstruction-based AD mix a pixel fidelity term with a structural term so that local structural dependencies are preserved on in-distribution data:

$$\mathcal{L}(\theta) = \|\mathbf{I} - g_\theta(\mathbf{I})\|_1 + \lambda\big(1 - \mathrm{SSIM}(\mathbf{I}, g_\theta(\mathbf{I}))\big), \tag{4.1}$$

with $\lambda \geq 0$ controlling the balance (Zhou and Paffenroth, 2017). There are a few common variants under this umbrella. *Masking and inpainting* methods hide random regions and learn to complete them from context, which encourages the model to rely on surrounding anatomy (Zavrtanik et al., 2020). *Transformer-based reconstructors* replace or augment convolutional operations to model longer-range dependencies in the image (You et al., 2022b). Finally, *diffusion-style reconstructors* denoise a corrupted version of the input toward a clean in-distribution-looking target (Teng et al., 2022). The core idea is the same: fit the in-distribution manifold well, so deviations stand out at test time.

After we compute $\hat{\mathbf{I}}$ (the reconstructed output image), we measure a per-pixel discrepancy. A simple choice is the element-wise absolute difference $|\,\mathbf{I} - \hat{\mathbf{I}}\,|$. Many works also use local structural dissimilarity, i.e., $(1 - \mathrm{SSIM})$, because it responds to texture changes and edge distortions that are typical for MR artifacts. A practical combination for the resulting anomaly map $\mathbf{M}$ is

$$\mathbf{M} = \alpha\,|\,\mathbf{I} - \hat{\mathbf{I}}\,| + (1-\alpha)\,\big(1 - \mathrm{SSIM}_{\mathrm{local}}(\mathbf{I}, \hat{\mathbf{I}})\big), \quad \alpha \in [0,1]. \tag{4.2}$$

Here, $\mathrm{SSIM}_{\mathrm{local}}$ refers to the Structural Similarity Index Measure computed over local patches (as opposed to globally). Feature-space comparisons are also used: pass $\mathbf{I}$ and $\hat{\mathbf{I}}$ through an encoder and take a distance (e.g., cosine or ℓ_2) between the corresponding feature vectors at each spatial location. In practice, the raw map is smoothed (small mean or Gaussian filter) and min–max normalized to stabilize the range. A binary mask is then obtained by thresholding the map, which serves to explicitly localize anomalous regions and facilitate slice-level classification. In practice, this threshold is typically selected by optimizing a performance metric, such as the F1-score, on a held-out validation set.

Because the signal comes from reconstruction failure, these methods are naturally sensitive to local texture and structure changes: motion streaks, metal-related signal voids, or strong bias fields. Masking/inpainting variants *perform well* at catching context violations (a region that cannot be plausibly filled from its neighborhood) (Bao et al., 2024). Transformer and diffusion reconstructors help when long-range consistency matters, but if the information bottleneck is not sufficiently tight, the model may *over-generalize* and partially "clean up" subtle anomalies. This trade-off is well noted in recent studies and benchmarks (Bao et al., 2024; You et al., 2022b; Teng et al., 2022).

4.1.1 DRAEM

Discriminatively trained Reconstruction Anomaly Embedding (DRAEM), proposed by Zavrtanik et al. (2021), uses two subnetworks that are trained together: a reconstructive inpainting network and a discriminative segmentation network. It first corrupts in-distribution images with synthetic anomalies and trains the reconstructor to bring them back to an in-distribution-looking state. In parallel, it trains a discriminator that processes the input and its reconstruction simultaneously, by concatenating them along the channel dimension, and learns to output a pixel-wise anomaly map (Zavrtanik et al., 2021). In this way, the model does not rely only on raw residuals and learns

a data-driven notion of "difference" between the two. Figure 4.1 illustrates the anomaly detection process of the DRAEM model.

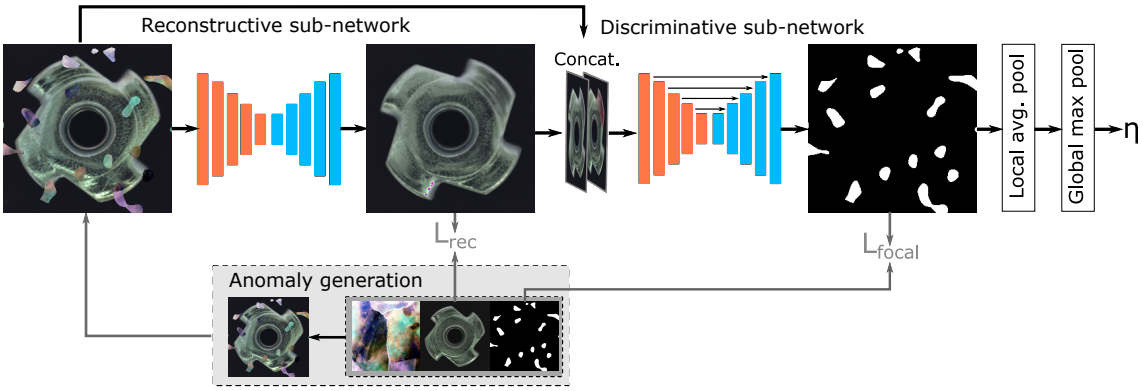


Figure 4.1: THE ANOMALY DETECTION PROCESS OF THE DRAEM MODEL. *Synthetic anomaly generation feeds a reconstructive sub-network (optimized via the reconstruction loss $\mathcal{L}_{rec}$) that outputs a reconstruction of the original image, while a U-Net-like (Ronneberger et al., 2015) discriminator (optimized via the focal loss $\mathcal{L}_{focal}$) takes the concatenation of the input and the reconstruction and predicts an anomaly map* $\mathbf{M}$ *(Zavrtanik et al., 2021).*

Formally, if $\mathbf{I}$ is an in-distribution image and $\mathbf{I}_a$ its corrupted version, the reconstructor g_θ produces $\mathbf{I}_r = g_\theta(\mathbf{I}_a)$. It is trained with a mixed fidelity/structure loss

$$\mathcal{L}_{\text{rec}}(\mathbf{I}, \mathbf{I}_r) \;=\; \lambda \left(1 - \text{SSIM}(\mathbf{I}, \mathbf{I}_r)\right) \;+\; \|\mathbf{I} - \mathbf{I}_r\|_2^2, \tag{4.3}$$

where the SSIM term keeps local contrast and edges stable and $\lambda > 0$ balances the two parts (Zavrtanik et al., 2021; Wang et al., 2004). The discriminator receives the channel-wise concatenation $\mathbf{I}_c = [\mathbf{I} \,\|\, \mathbf{I}_r]$ and outputs a per-pixel anomaly probability map $\mathbf{M} \in [0,1]^{H \times W}$, where H and W denote the image height and width, respectively. Supervision comes from the known synthetic mask $\mathbf{M}_a$ using a segmentation loss (implemented as focal loss to reduce the impact of easy negatives). As shown in Figure 4.1, the total objective combines the reconstruction loss ($\mathcal{L}_{\text{rec}}$) and the discriminative focal loss ($\mathcal{L}_{\text{focal}}$):

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{rec}}(\mathbf{I}, \mathbf{I}_r) \;+\; \mathcal{L}_{\text{focal}}(\mathbf{M}, \mathbf{M}_a). \tag{4.4}$$

The synthetic anomalies are designed to be varied at the texture level. DRAEM samples organic-shaped masks—meaning masks with irregular, non-geometric boundaries—using thresholded Perlin noise and fills them with foreign textures drawn from an external source (Perlin, 1985). Blending uses a random opacity $\beta \in [0.1, 1.0]$ (Zavrtanik et al., 2021):

$$\mathbf{I}_a \;=\; (1 - \mathbf{M}_a) \odot \mathbf{I} \;+\; (1 - \beta)\left(\mathbf{M}_a \odot \mathbf{I}\right) \;+\; \beta\left(\mathbf{M}_a \odot \mathbf{A}\right), \tag{4.5}$$

where $\odot$ denotes the element-wise product and $\mathbf{A}$ the anomaly texture. This procedure exposes the networks to a wide range of shapes, contrasts, and boundaries during training.

During inference, the synthetic anomaly generation process is omitted. The unmodified test image is passed directly through the reconstructive sub-network to generate the reconstruction $\mathbf{I}_r$. Subsequently, the discriminative sub-network processes the concatenation of the original test image and $\mathbf{I}_r$ to predict the anomaly map $\mathbf{M}$. For a single slice-level score, the map is smoothed to

aggregate local anomaly responses. This is performed by convolving the map with a mean filter $\mathbf{F}_{s\times s}$ before taking the maximum value:

$$\eta \;=\; \max\big(\mathbf{M} * \mathbf{F}_{s\times s}\big), \tag{4.6}$$

where $*$ denotes the convolution operation. The kernel $\mathbf{F}_{s\times s}$ is defined as an $s \times s$ matrix where every element is equal to $1/s^2$, acting as a constant averaging filter (Zavrtanik et al., 2021). This summarizes how strongly any region in the image looks anomalous. In practice, the discriminator is implemented as a U-Net-like segmenter (Ronneberger et al., 2015). Using patchwise SSIM in the reconstruction loss helps the reconstructor keep small structures sharp, and focal loss stabilizes the segmentation under the strong class imbalance that comes from sparse anomaly masks (Lin et al., 2017).

4.1.2 Dinomaly

Dinomaly is a reconstruction-based AD model that keeps the design minimal and builds everything from basic Transformer blocks (Dosovitskiy et al., 2021). It targets the "identity mapping" pitfall-a phenomenon where the network learns to trivially copy input features to the output, thereby reconstructing out-of-distribution patterns instead of exhibiting the desired reconstruction error (Guo et al., 2025). To mitigate this, the architecture incorporates three specific components: a noisy bottleneck to discourage trivial passthrough (You et al., 2022a), a decoder with linear attention to prevent the model from focusing on exact spatial locations, and a *loose* reconstruction target that supervises grouped features rather than enforcing strict layer-to-layer correspondence (Guo et al., 2025). The aim is to reconstruct in-distribution *features* accurately without copying the input pixel-by-pixel, ensuring that out-of-distribution regions manifest as feature mismatches.

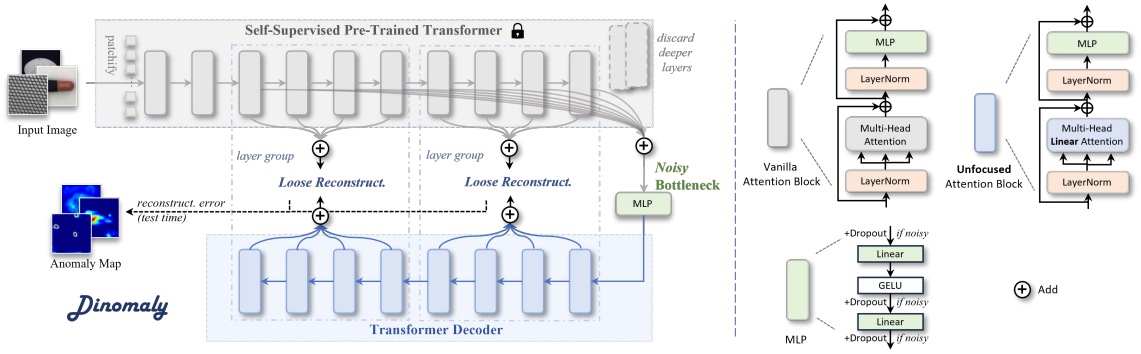


Figure 4.2: THE FRAMEWORK OF DINAMOLY IS CONSTRUCTED USING BASIC TRANSFORMER BUILDING BLOCKS. *A ViT-style encoder produces grouped features; a noisy bottleneck and a transformer decoder with linear attention reconstruct mid-level features (Zhang et al., 2024); the feature mismatch forms the anomaly map (Guo et al., 2025).*

As depicted in Figure 4.2, given an input image $\mathbf{I}$, an encoder f_E outputs multi-layer features $\{f_E^\ell(\mathbf{I})\}_{\ell=1}^L$. These features are passed through a lightweight bottleneck with Dropout to add controlled noise, producing tokens that serve as the input to the decoder. The decoder f_D then reconstructs mid-level features. Within a decoder block, the input token sequence X is projected

to form queries (Q), keys (K), and values (V):

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V. \tag{4.7}$$

where W_Q, W_K, and W_V are the learnable weight matrices for the query, key, and value projections, respectively. Instead of the conventional softmax attention,

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V, \tag{4.8}$$

where d is the dimension of the key vectors, the decoder uses linear attention to reduce sharp, location-specific copying (Katharopoulos et al., 2020):

$$\mathrm{LA}(Q, K, V) = \phi(Q)\big(\phi(K)^\top V\big), \qquad \phi(x) = \mathrm{elu}(x) + 1. \tag{4.9}$$

Instead of enforcing a strict layer-to-layer reconstruction constraint, Dinomaly aggregates encoder layers into a few hierarchical groups (e.g., low and mid). For a group g with index set $\mathcal{G}_g$, encoder features are aggregated via element-wise summation, $\mathbf{h}_g = \sum_{\ell \in \mathcal{G}_g} f_E^\ell(\mathbf{I})$. This operation relies on the columnar structure of the Transformer backbone, which ensures that feature maps across distinct layers maintain identical spatial dimensions and channel counts. The corresponding decoder output $\hat{\mathbf{h}}_g$ is then taken as the reconstruction target. The training objective is a global cosine loss over groups, with hard mining to down-weight already well-reconstructed points (Guo et al., 2025):

$$\mathcal{L}_{\text{global}} = \frac{1}{G}\sum_{g=1}^{G}\Big(1 - \frac{\langle \mathrm{F}(\mathbf{h}_g), \mathrm{F}(\hat{\mathbf{h}}_g)\rangle}{\|\mathrm{F}(\mathbf{h}_g)\|\,\|\mathrm{F}(\hat{\mathbf{h}}_g)\|}\Big), \tag{4.10}$$

where G is the total number of feature groups, $\mathrm{F}(\cdot)$ flattens feature maps, $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\|\cdot\|$ is the L2 norm.

At test time, anomaly evidence is computed in feature space. For each spatial location $\mathbf{u}$ and group g, the cosine distance between encoder and decoder features is calculated and averaged across groups to obtain an anomaly score map $\mathbf{S}_{AL}$:

$$\mathbf{S}_{AL}(\mathbf{u}) = \frac{1}{G}\sum_{g=1}^{G}\Big(1 - \frac{\langle \mathbf{h}_g(\mathbf{u}), \hat{\mathbf{h}}_g(\mathbf{u})\rangle}{\|\mathbf{h}_g(\mathbf{u})\|\,\|\hat{\mathbf{h}}_g(\mathbf{u})\|}\Big) \tag{4.11}$$

where $\mathbf{h}_g(\mathbf{u})$ and $\hat{\mathbf{h}}_g(\mathbf{u})$ represent the feature vectors from the encoder and decoder, respectively, at spatial location $\mathbf{u}$ for group g.

In practice, the noisy bottleneck and linear attention reduce exact copying. This is complemented by the grouped, loose target, which keeps supervision *global* rather than pixel-accurate.

4.2 One-Class Classification Methods

In contrast to reconstruction-based methods, which aim to reproduce the input appearance, One-Class Classification (OCC) methods adopt a projection-based strategy. These methods map in-distribution samples into a compact region of feature space without explicit supervision, such that deviations from this region can be interpreted as anomalies. According to the taxonomy presented in Benchmarks for Medical Anomaly Detection from BMAD (Bao et al., 2024), two representative approaches within this category are Deep Support Vector Data Description (DeepSVDD) (Ruff et al., 2018) and CutPaste (Li et al., 2021a).

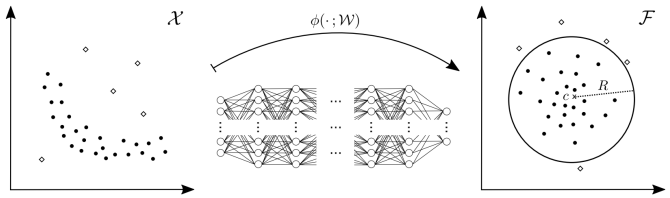


Figure 4.3: PRINCIPLE OF DEEPSVDD. *The network $\phi(\cdot; W)$ with weights W projects inputs into a latent feature space in which in-distribution samples cluster around a predefined center c. Anomalies are detected based on their distance from this center (Ruff et al., 2018).*

Classical OCC methods are based on principles such as the One-Class Support Vector Machine (OCSVM) (Schölkopf et al., 1999) and SVDD (Tax and Duin, 2004), which define a minimum-volume hypersphere that encloses in-distribution data. Deep variants, including DeepSVDD, extend this concept by employing an end-to-end objective that drives the network to compress in-distribution representations toward a compact latent center (Chalapathy and Chawla, 2019; Ruff et al., 2018, 2021).

Another line of OCC research uses self-supervised auxiliary tasks to learn feature spaces in which most data form tight, discriminative clusters. In this approach, as exemplified by CutPaste, synthetic perturbations are introduced during training not to approximate the distribution of true anomalies, but to provide a contrastive signal that encourages the model to learn stable representations of in-distribution structure. Importantly, inference still relies solely on evaluating how far a sample is from the learned representation of in-distribution data, thereby preserving the task's one-class nature.

OCC methods primarily generate sample-level anomaly scores that inform image- or slice-level decisions. While some variants enable coarse localization through post-hoc interpretation techniques, they do not inherently produce dense pixel-level anomaly maps as reconstruction-based methods do. Methods such as DeepSVDD and CutPaste remain widely adopted baselines for out-of-distribution detection in medical imaging benchmarks (Cao et al., 2020).

4.2.1 DeepSVDD

DeepSVDD (Ruff et al., 2018) adapts the classical SVDD principle to the deep learning context. A fully deep approach streamlines the detection pipeline by bypassing pixel-wise reconstruction entirely. The core mechanism involves training a neural encoder $f_\theta(\cdot)$ to map all in-distribution training images into a minimum-volume hypersphere centered at c in the latent space (see Figure 4.3). This process minimizes the variance of the in-distribution data around the center:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|f_\theta(\mathbf{I}_i) - c\|_2^2 \tag{4.12}$$

To avoid the trivial solution in which all embeddings collapse to a single point, the center c is not considered a trainable parameter. Consistent with established methodology (Ruff et al., 2018), c is determined by computing the mean of the embeddings generated during an initial forward pass over the training data with the network's initial weights.

At inference, the anomaly score $s(\mathbf{I})$ is defined as the squared Euclidean distance between the embedding and the center:

$$s(\mathbf{I}) = \|f_\theta(\mathbf{I}) - c\|_2^2. \tag{4.13}$$

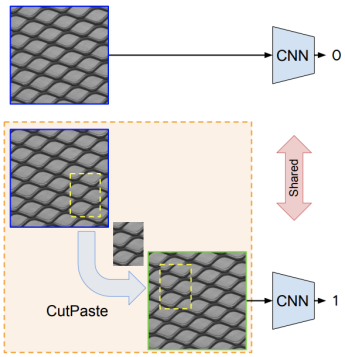


Figure 4.4: CUTPASTE REPRESENTATION LEARNING FOR ANOMALY DETECTION.. *A representation is trained to separate original in-distribution images from synthetically augmented variants generated by cutting and pasting a patch at a random location (Li et al., 2021a).*

When the encoder uses global average pooling, spatial feature maps are compressed into a single vector before projection. This design enables slice-level detection but prevents the model from producing pixel-wise anomaly maps.

4.2.2 CutPaste

CutPaste (Li et al., 2021a) reformulates one-class classification as a self-supervised learning problem by generating synthetic perturbations and training the network to differentiate these from in-distribution samples. Figure 4.4 demonstrates the construction of such proxy anomalies. In contrast to DeepSVDD, which enforces geometric compactness, CutPaste uses a contrastive training signal to shape the embedding space. The cut-and-paste operations create texture irregularities and boundary discontinuities that disrupt the local statistical regularities of anatomical structures. These perturbations are applied exclusively during training to provide a learning signal. During inference, anomaly detection is based solely on deviations from the learned in-distribution representation, thereby maintaining the one-class paradigm.

During training, the image encoder is optimized with a self-supervised classification objective that distinguishes original images from multiple CutPaste-based augmented variants. This objective can be formulated using a standard cross-entropy loss over the constructed proxy classes,

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k} y_{i,k}\log p_{i,k}, \tag{4.14}$$

where $p_{i,k}$ represents the predicted probability for proxy class k. This process enables the encoder to learn an embedding in which in-distribution samples cluster together, while perturbed samples are separated.

Following training, anomaly scoring is performed using Gaussian Density Estimation (GDE) on the learned embeddings. A multivariate Gaussian distribution is fitted to the embeddings of the in-distribution training data, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The anomaly score for a test image s is computed as the log-density of its embedding $f(s)$ under this distribution:

$$\log p_{\text{GDE}}(s) \propto \{-\frac{1}{2}(f(s)-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(f(s)-\boldsymbol{\mu})\}. \tag{4.15}$$

A lower likelihood indicates a greater deviation from the nominal data distribution, resulting in a higher anomaly score. Since the encoder generates a single embedding per image, this density-based approach inherently produces slice-level anomaly scores.

4.3 Knowledge Distillation Methods

Knowledge Distillation (KD) is a model compression and transfer technique where the student model accepts raw data/images as input and attempts to match the output of the pre-trained teacher model (Hinton et al., 2015). Modern approaches extend this, focusing also on aligning intermediate feature representations between teacher and student, a process known as feature-level distillation (Romero et al., 2015). In unsupervised AD, KD is consequently repurposed as a normality estimator (Bergmann et al., 2020).

Residual Networks (ResNet) proposed by He et al. (2016), are a widely used family of deep convolutional neural networks that leverage skip connections to facilitate the training of very deep architectures, enabling them to achieve strong performance on various challenging visual tasks. The ImageNet dataset, introduced by Deng et al. (2009), containing millions of natural images across thousands of categories, is commonly used for pretraining visual models and is regarded as a benchmark for general visual representation learning.

Building on these foundations, the KD models considered in this work instantiate the teacher as a large, frozen ResNet backbone pretrained on ImageNet, which serves as a robust expert feature extractor for the student (Deng and Li, 2022; Wang et al., 2021). This teacher serves as a robust expert in extracting generic features from diverse images without prior exposure to anomalies in the target domain. The student model is trained only on in-distribution data from the target domain to replicate the teacher's features for these patterns. The underlying assumption is that such a student will closely match the teacher on in-distribution inputs but tend to deviate on anomalous ones, leading to a larger discrepancy between their feature representations (Bergmann et al., 2020). In practice, this discrepancy is measured using distance or similarity functions between teacher and student feature maps and aggregated into an anomaly score.

This KD paradigm is highly advantageous for medical image AD, especially under conditions of data scarcity and diversity. In particular, it leverages the rich, general knowledge encoded within the pre-trained teacher model, which is beneficial for domains with limited annotated data (Salehi et al., 2021). Additionally, KD-based models allow anomaly segmentation by comparing feature maps at multiple scales. Discrepancies between teacher and student feature representations can then highlight atypical image regions, aiding identification of anomalous areas in medical images (Deng and Li, 2022; Wang et al., 2021). Because teacher and student architectures can differ in depth and width, internal features are not directly comparable in general. In practice, feature-based KD methods therefore either choose architecturally compatible layers or introduce lightweight projection modules so that corresponding feature maps can be aligned and compared (Deng and Li, 2022; Wang et al., 2021). These alignment mechanisms are optimized jointly with the student during training so that the internal representations approximate those of the teacher on training data (Deng and Li, 2022). The concrete instantiations of this process are architecture-dependent and are detailed for each model in its respective section.

Several models were developed that make use of the knowledge distillation paradigm, of which 2 are explored in this work. STFPM was proposed by Wang et al. (2021) and leverages a multi-scale feature discrepancy between teacher and student for AD while keeping the architecture of both similar. RD4AD, proposed by Deng and Li (2022), introduced the paradigm of reverse distillation with an architecture that uses a teacher Encoder and student Decoder, explicitly enhancing the feature discrepancy on OOD samples through architectural design.

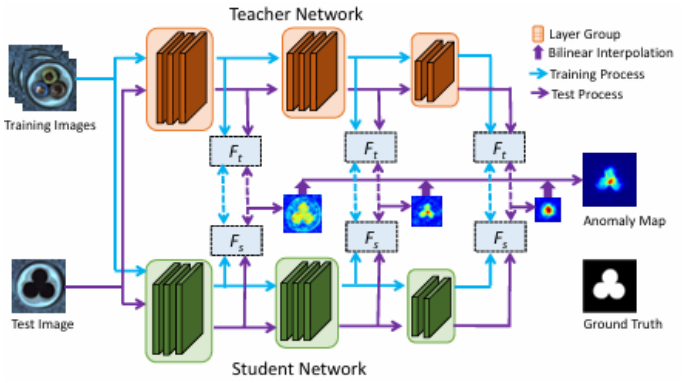


Figure 4.5: SCHEMATIC OVERVIEW OF THE STFPM METHOD. *The figure provides an overview of how multi-layer feature maps from the student are trained to match their counterparts in the teacher network through Feature Pyramid Matching (Wang et al., 2021).*

4.3.1 STFPM - Student-Teacher Feature Pyramid Matching

Student-Teacher Feature Pyramid Matching (STFPM), proposed by Wang et al. (2021), is a straightforward KD-based approach designed for pixel-level anomaly detection and localization. STFPM enhances knowledge transfer through hierarchical feature matching.

STFPM adopts the standard KD paradigm and introduces a key architectural choice by using identical architectures for both the teacher and student networks. The teacher is implemented as a frozen, pre-trained backbone, such as ResNet-18 pretrained on ImageNet. The student is a trainable counterpart with the same architecture. This mirroring of architecture allows knowledge transfer in a single step from the teacher to the student, and maximizes the retention of crucial feature information. As a result, STFPM addresses the challenge of incomplete knowledge transfer that is often encountered when employing structurally different student models in other KD approaches.

The defining aspect of STFPM is the use of hierarchical, multi-layer feature matching between teacher and student. Visualization studies have shown that convolutional networks form hierarchical feature representations, where early layers respond to edge- and texture-like patterns and deeper layers capture more abstract object parts and structures (Zeiler and Fergus, 2014). The Feature Pyramid Matching (FPM) block aggregates features from several successive layers of the teacher to guide the student, so that information from different semantic levels is jointly aligned. This hierarchical matching enables the model to detect and localize anomalies across a range of sizes. Figure 4.5 presents an overview of how the multi-layer feature maps are used to create the final anomaly map.

During training, only anomaly-free images are used. The student network is updated while the teacher remains fixed. The goal is for the student's multi-layer feature representations to closely imitate those of the teacher at all selected layers.

The loss at a specific layer l and position (h, w) where $h \in \{1, \ldots, H\}$ and $w \in \{1, \ldots, W\}$ is defined as the ℓ_2-distance between the ℓ_2-normalized feature vectors of the teacher and the student. Let $\mathbf{f}_l^t(h, w)$ and $\mathbf{f}_l^s(h, w)$ be the normalized feature vectors at layer l and position (h, w) from the teacher and student, respectively. The per-pixel loss $\mathcal{L}_l(h, w)$ at this position is:

$$\mathbf{M}_l(h, w) = \frac{1}{2}||\mathbf{f}_l^t(h, w) - \mathbf{f}_l^s(h, w)||_{l_2}^2 \tag{4.16}$$

Since the feature vectors are ℓ_2-normalized, this ℓ_2-distance is proportional to the cosine distance.

The total loss $\mathcal{L}_{KD}$ for an image is the weighted average of the loss across all l selected pyramid layers, where the loss for each layer $\mathcal{L}_l$ is the average of the per-pixel losses at that layer:

$$\mathcal{L}_{KD} = \sum_{l=1}^{L} \alpha_l \cdot \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \mathbf{M}_l(h, w) \tag{4.17}$$

In Equation (4.17), the weight α_l is introduced to allow different feature scales to be emphasized if desired. However, in STFPM and in this work, all scales are treated equally by setting $\alpha_l = 1$. Minimizing the loss $\mathcal{L}_{KD}$ ensures the student accurately learns the in-distribution features represented by the teacher.

At inference time, the model computes an anomaly map $\mathbf{M}_l$ for each of the L selected layers using the discrepancy defined in Equation (4.16). To create the final, full-resolution anomaly score map ($\mathbf{S}_{AL}$), the individual feature maps are upsampled using bilinear interpolation (Ψ), such that they have the same dimensions, and then combined through an element-wise product:

$$\mathbf{S}_{AL} = \prod_{l=1}^{L} \Psi(\mathbf{M}_l) \tag{4.18}$$

The use of the element-wise product ensures that a high anomaly score is assigned only to regions where the discrepancy is high across multiple feature scales, leading to precise localization.

4.3.2 RD4AD - Reverse Distillation for Anomaly Detection

The Reverse Distillation for Anomaly Detection (RD4AD) model, proposed by Deng and Li (2022), introduces a distinct strategy to KD to enhance the detection and localization of OOD samples. Traditional KD models for anomaly detection often suffer from a vanishing feature discrepancy between the teacher and student networks when encountering anomalies, a failure often attributed to the similarity in their network architectures and data flow. To address this, RD4AD proposes a Reverse Distillation paradigm that changes the architecture. Specifically, the teacher network is implemented as a high-capacity encoder. This deep encoder extracts multi-scale feature representations and serves as the expert for identifying in-distribution image patterns within the target domain. The student network is constructed as a trainable decoder with an architecture that mirrors but reverses the structure of the teacher, enabling effective reconstruction of the teacher's encoded features from compressed representations. The student learns to reconstruct anomaly-free feature representations from the compressed knowledge passed by the teacher. This reconstruction task is more sensitive to structural completeness, forcing the student to fail distinctly when presented with OOD information, thereby maximizing the feature discrepancy for robust OOD detection.

An important component of the RD4AD network is the trainable One-Class Bottleneck Embedding (OCBE) module, which connects the teacher Encoder to the student Decoder. The OCBE serves to both compress the teacher's knowledge and actively suppress anomalous information, making the student a more robust anomaly-rejector. The OCBE achieves this through two integrated parts: the Multi-Scale Feature Fusion (MFF) block, which aggregates features from both shallow (low-level details) and deep (high-level structural and semantic information) layers, and dimensionality reduction, where the aggregated features are projected into a compact code. This bottleneck acts as a filter that retains essential features of in-distribution patterns but is too restrictive to propagate the specific perturbations that correspond to anomalies, ensuring the student receives an anomaly-filtered input. Figure 4.6 shows an overview of the RD4AD network and how each component is connected to each other, creating the encoder-decoder structure.

During training, only in-distribution images are used. The student and the OCBE module are jointly optimized to minimize the knowledge distillation loss ($\mathcal{L}_{KD}$), which is calculated as the

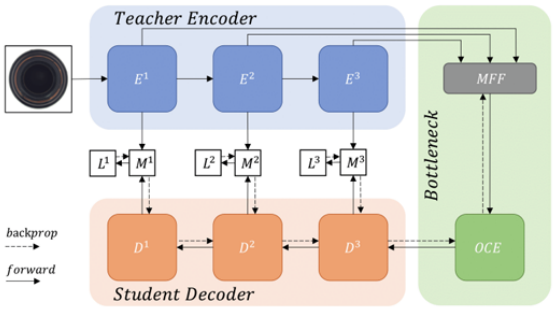


Figure 4.6: SCHEMATIC OVERVIEW OF THE RD4AD METHOD. *The figure shows an overview of how the RD4AD network follows an encoder–decoder structure and depicts the comparison of multi-scale feature embeddings between encoder and decoder (Deng and Li, 2022).*

accumulation of pixel-wise cosine similarity losses across L feature layers. Let $\mathbf{f}_l^E$ and $\mathbf{f}_l^D$ be the feature tensors from the l-th encoding and decoding blocks, respectively. At each spatial location (h, w), we take the channel vectors $\mathbf{f}_l^E(h, w)$, $\mathbf{f}_l^D(h, w)$ and define the anomaly map $\mathbf{M}_l$ as:

$$\mathbf{M}_l(h, w) = 1 - \frac{(\mathbf{f}_l^E(h, w))^T \cdot \mathbf{f}_l^D(h, w)}{||\mathbf{f}_l^E(h, w)|| \cdot ||\mathbf{f}_l^D(h, w)||} \tag{4.19}$$

The final loss $\mathcal{L}_{KD}$ is defined analogously to STFPM in (4.17) as the accumulated mean of the multi-scale anomaly maps. Minimizing this loss forces the Decoder to closely reproduce the Encoder's features for in-distribution patterns.

At inference, the model receives an input image and generates an Anomaly Score Map ($\mathbf{S}_{AL}$) by calculating the feature discrepancy using (4.19) across all L selected layers and accumulating the upsampled maps.

$$\mathbf{S}_{AL} = \sum_{l=1}^{L} \Psi(\mathbf{M}_l) \tag{4.20}$$

Unlike STFPM, which combines per-layer anomaly maps via an element-wise product (4.18), RD4AD aggregates the upsampled maps by summation (4.20). Summation treats each scale as contributing additively to the anomaly evidence, allowing a strong signal at a single scale to produce a high anomaly score. In contrast, the multiplicative aggregation in STFPM emphasizes regions that are consistently anomalous across scales, potentially suppressing anomalies that appear prominently only at one feature level.

The $\mathbf{S}_{AL}$ provides precise localization by assigning higher values to regions where the student, having only learned in-distribution patterns, failed to mimic the teacher's truthful output. For slice-level anomaly detection, the final anomaly score is derived as the maximum value in the score map to ensure sensitivity to small anomalous regions.

4.4 Memory Bank

Memory bank methods (Bao et al., 2024; Defard et al., 2021; Lee et al., 2022; Roth et al., 2022; Li et al., 2021b), also called exemplar-based methods, are characterized by explicitly storing feature prototypes of in-distribution data. During inference, each feature from a test image is compared

to this memory bank. Large feature-space deviations from all stored prototypes indicate OOD regions.

In practice, this comparison can be implemented via distance computations or statistical modeling: for example, k-NN voting (Prerau and Eskin, 2000; Roth et al., 2022), or Mahalanobis distances against Gaussian models of the memory (Defard et al., 2021). In this way, a test image is scored based on how far its features lie from the learned in-distribution manifold, rather than how poorly it can be reconstructed or how much a student network deviates from a teacher.

Memory-bank models typically operate on pretrained feature maps (often multi-scale CNN embeddings) extracted from frozen backbones. At test time, they require distance or similarity computations between test features and stored prototypes. While some methods use simple nearest-neighbor distances, others employ parametric statistical modeling such as fitting multivariate Gaussians, which involves estimating mean vectors and covariance matrices from the training data. Prominent examples include PatchCore and CFA. PatchCore (Roth et al., 2022), which grounds nearest-neighbor anomaly detection in principles similar to those explored in extreme value machine approaches (Rudd et al., 2018), compiles a coreset of in-distribution patch features into a memory bank and assigns each test patch an anomaly score equal to its distance to the nearest memory neighbor. CFA (Lee et al., 2022) goes further by training a lightweight patch-descriptor network on in-distribution data so that its embeddings form tight hyperspherical clusters; the centers of these clusters populate the memory, and anomaly scores are given by nearest-cluster distances.

By leveraging fixed feature extractors and nearest-neighbor or probabilistic scoring, memory-bank methods often achieve efficient inference with interpretable, localized anomaly outputs. In the context of MR-only radiotherapy, they combine powerful pretrained image features with distance-based anomaly metrics. In effect, memory-bank approaches bridge the gap between feature-domain and pixel-domain strategies. This hybrid nature makes them a valuable addition to the anomaly-detection toolkit for medical imaging, providing efficient and explainable localization of OODs.

4.4.1 CFA

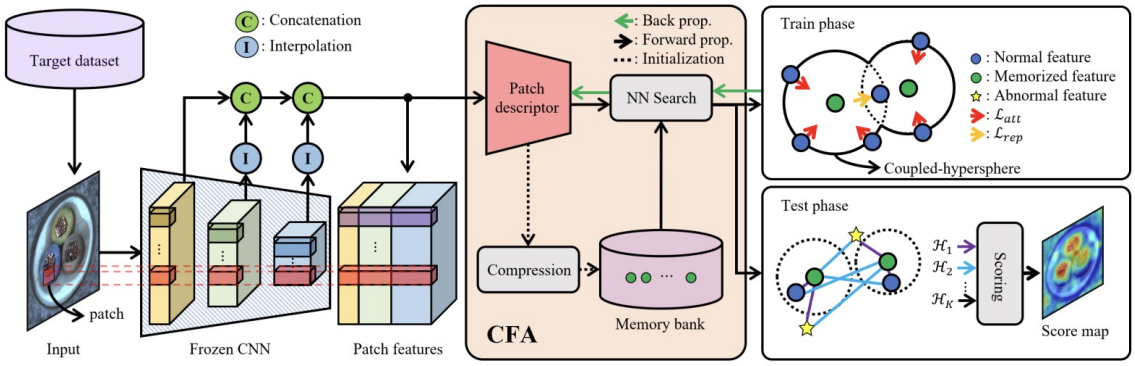


Figure 4.7: OVERVIEW OF CFA METHOD. *CFA trains a frozen CNN's patch descriptor via coupled-hypersphere supervision ($\mathcal{L}_{att}$, $\mathcal{L}_{rep}$) against compressed in-distribution prototypes. At test time, NN search matches patches; OOD features are scored via certainty-aware function to localize anomalies.*

CFA adapts the feature space of in-distribution images into the target domain by learning a domain-specific embedding space to minimize the bias of pre-trained CNN features. Given a set of N in-distribution training images, each image $\mathbf{I}$ is divided into non-overlapping spatial patches. Multi-scale patch maps are extracted from intermediate layers of a pre-trained CNN backbone $\phi(\cdot)$. For each spatial position (h, w) on this feature map, let $\mathbf{f}_{(h,w)} \in \mathbb{R}^D$ denote the raw patch feature extracted, where D is the feature dimension of the pre-trained backbone. The trainable patch descriptor network $\psi(\cdot)$ then maps each patch into an adapted feature space:

$$\mathbf{v}_{(h,w)} = \psi(\mathbf{f}_{(h,w)}) \in \mathbb{R}^d$$

where d is the dimension of the learned embedding space.

During training, CFA constructs a memory bank $C = \{c^{(1)}, c^{(2)}, \ldots, c^{(M)}\}$ containing M prototype features from the in-distribution training set. For each adapted patch feature $\mathbf{v}_{(h,w)}$, CFA identifies its K nearest neighbors $\{c^{(k)}_{(h,w)}\}_{k=1}^{K} \subset C$ in the memory bank, where neighbors are indexed by their ranking (k-th nearest; superscripts denote ranking). The training objective enforces that $\mathbf{v}_{(h,w)}$ should lie within a hypersphere of radius r centered on these neighbors. The attractive loss $\mathcal{L}_{\text{att}}$ penalizes embedded features that lie farther than r from all of their K-th nearest neighbors:

$$\mathcal{L}_{\text{att}} = \frac{1}{HWK} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{k=1}^{K} \max \left\{ 0, \left\| \mathbf{v}_{(h,w)} - c^{(k)}_{(h,w)} \right\|_2^2 - r^2 \right\}, \tag{4.21}$$

This term encourages adapted patch features to form compact clusters within the in-distribution feature manifold.

However, features lying within multiple overlapping hyperspheres may cause OOD regions to be misclassified as in-distribution. To prevent this, CFA introduces hard negative features, defined as the (K+1)-th through (K+J)-th nearest neighbors $\{c^{j}_{(h,w)}\}_{j=1}^{J}$. The repulsive loss $\mathcal{L}_{\text{rep}}$ applies contrastive supervision to push $\mathbf{v}_{(h,w)}$ away from these hard negatives:

$$\mathcal{L}_{\text{rep}} = \frac{1}{HWJ} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{j=1}^{J} \max \left\{ 0, r^2 - \left\| \mathbf{v}_{(h,w)} - c^{(j)}_{(h,w)} \right\|_2^2 - \alpha \right\}, \tag{4.22}$$

where α is a margin hyperparameter used to control the balance between $\mathcal{L}_{att}$ and $\mathcal{L}_{rep}$. , J is the number of hard negative features $\{c_t^{(j)}\}_{j=1}^{J}$ defined as the next nearest neighbors beyond K.

The total training objective combines both effects:

$$\mathcal{L}_{\text{CFA}} = \mathcal{L}_{\text{att}} + \mathcal{L}_{\text{rep}} \tag{4.23}$$

Optimizing $\phi(\cdot)$ with $\mathcal{L}_{\text{CFA}}$ adapts the pre-trained features to form compact clusters around in-distribution prototypes through attraction, and creates clear decision boundaries that separate in-distribution from OOD regions through repulsion with margin.

During inference on a test image $\mathbf{I}_{\text{test}}$, CFA computes anomaly scores on the feature-map grid. The local anomaly score is defined as the squared Euclidean distance to the nearest nominal feature:

$$S_{AL}(h, w) = \min_{k \in \{1,\ldots,K\}} \left\| \mathbf{v}_{(h,w)} - c^{(k)}_{(h,w)} \right\|_2^2. \tag{4.24}$$

To account for the confidence of the nearest-neighbor match, CFA introduces a confidence weighted anomaly score

$$A(h, w) = \frac{\exp(-S_{AL}(h, w))}{\sum_{k=1}^{K} \exp\left(- \left\| \mathbf{v}_{(h,w)} - c^{(k)}_{(h,w)} \right\|_2^2 \right)} \cdot S_{AL}(h, w). \tag{4.25}$$

The resulting anomaly scores form a feature-space anomaly map. For pixel-level localization, the anomaly score is upsampled to the original image resolution using bilinear interpolation, and subsequently smoothed using Gaussian filtering to obtain the final anomaly localization map. Regions with higher anomaly scores indicate greater deviation from the learned in-distribution feature manifold.

CFA effectively reduces the domain bias of pre-trained CNNs by learning a compact, domain-specific feature space that tightly clusters in-distribution features while separating them from anomalous ones. This adaptive representation improves both anomaly detection accuracy and localization precision, producing detection accuracy and localization precision, yielding reliable and well-localized anomaly maps.

4.4.2 PatchCore

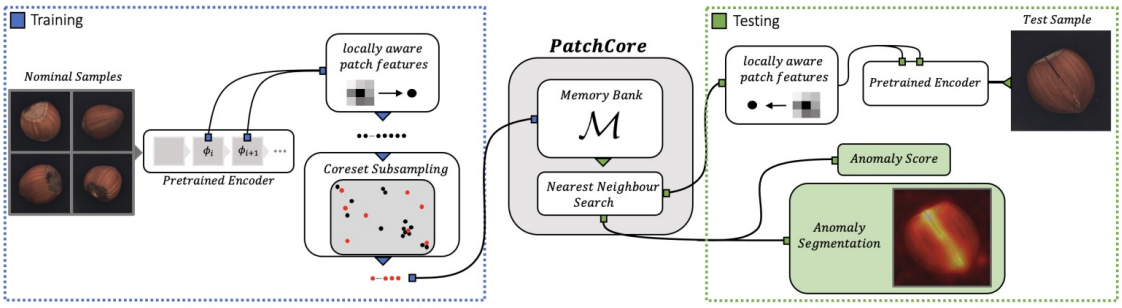


Figure 4.8: OVERVIEW OF PATCHCORE METHOD. *PatchCore builds a coreset-subsampled bank of locally-aware nominal patch features. Nearest neighbor search at test time yields anomaly scores for detection and localization.*

For an input image $\mathbf{I}$ from the training set, the feature map at layer k is denoted $f_k(I) \in \mathbb{R}^{C_k \times H_k \times W_k}$, where C_k is the number of channels and $H_k \times W_k$ is the spatial resolution of the feature map.

Each spatial feature vector corresponds to a receptive field in the input image and can therefore be interpreted as a patch descriptor. To increase the effective receptive field while retaining spatial resolution, PatchCore aggregates features over a local neighborhood.

Let n_s denote the neighborhood size. The neighborhood centered at (h, w) is defined as

$$N_{(h,w)}^{n_s} = \{(a, b) \mid a \in [h - \lfloor n_s/2 \rfloor, h + \lfloor n_s/2 \rfloor],\, b \in [w - \lfloor n_s/2 \rfloor, w + \lfloor n_s/2 \rfloor]\}, \tag{4.26}$$

$$f_k(N_{(h,w)}^{n_s}) = f_{\text{agg}}\left(\{f_k^{(a,b)} \mid (a, b) \in N_{(h,w)}^{n_s}\}\right), \tag{4.27}$$

where f_{agg} denotes average pooling over the fixed neighborhood.

The patch feature collection with optional spatial stride s is defined as

$$P_{s,p}(f_k = \left\{ f_k(N_{(h,w)}^{n_s}) \,\middle|\, h, w \in \mathbb{N},\, h < H_k, w < W_k,\, h \bmod s = 0, w \bmod s = 0 \right\}, \tag{4.28}$$

When $s = 1$, patch features are extracted densely at every spatial location, preserving the full feature-map resolution. Larger strides reduce redundancy at the cost of spatial granularity.

To incorporate multiscale context, PatchCore uses features from two adjacent hierarchy levels k and $k+1$. Because deeper layers have lower spatial resolution, the feature map at level $k+1$ is upsampled via bilinear to level k. The full memory bank M of nominal features is constructed by aggregating all patch features across the training dataset $X_N = \{I_1, \ldots, I_N\}$:

$$\mathcal{M} = \bigcup_{I \in X_N} P_{s,p}(f_k(\mathbf{I})). \tag{4.29}$$

To improve efficiency, PatchCore applies coreset sampling to select a representative subset $M_K \subset M$ using a greedy minimax facility location algorithm, reducing computational cost while maintaining coverage.

At inference, the slice-level anomaly score is defined as:

$$s = \max_{m_{\text{test}} \in P_{s,p}(I_{\text{test}})} \min_{m \in \mathcal{M}_K} \|m_{\text{test}} - m\|_2, \tag{4.30}$$

Let $m_{\text{test}}^{\max}$ denote the test patch attaining the maximum in the above expression, and let

$$m^{(1)}, \ldots, m^{(K)} \in \mathcal{M}_K$$

The slice-level score is reweighted by a softmax-based contrast that reflects the rarity of the closest match:

$$s \leftarrow \left(1 - \frac{\exp(\|m_{\text{test}}^{max} - m^{(1)}\|_2)}{\sum_{m \in C} \exp(\|m_{\text{test}}^{max} - m^{(k)}\|_2)}\right) \cdot s, \tag{4.31}$$

Patch-level anomaly scores are projected back to their corresponding spatial locations to obtain a dense anomaly map. Similar to CFA, it is followed by upsampling and Gaussian smoothing steps.

PatchCore performs anomaly detection comparing each test patch directly with a set of representative patch-level neighbors drawn from the training data. Through greedy coreset sampling, it retains diverse yet compact reference patches, allowing for efficient nearest-neighbor search during inference. This patch-wise matching enables the model to detect subtle, localized anomalies by leveraging actual examples of normality, providing a balance between detection accuracy, interpretability, and computational efficiency.

4.5 Flow-Based Methods

According to Kobyzev et al. (2021), flow-based methods are a class of generative models that allow for explicit likelihood estimation through a series of invertible transformations. Compared to reconstruction-based approaches such as autoencoders or diffusion models, which rely on implicit density estimation, normalizing flows (NFs) enable the exact computation of the likelihood of an input image. This provides a well-defined measure for how likely an image is under a learned distribution rather than relying on an approximate measure like the reconstruction error.

A normalizing flow defines a bijective mapping $f_\theta : \mathcal{I} \to \mathcal{Z}$ that transforms an input image $\mathbf{I}$ into a latent variable $z \in \mathbb{R}^d$ followed by a simple base distribution $p_Z(Z)$, typically a standard normal $\mathcal{N}(0, \mathbf{I}_d)$. This allows us to express the likelihood of $\mathbf{I}$ to as:

$$\log p_{\mathbf{I}}(\mathbf{I}) = \log p_Z(f_\theta(\mathbf{I})) + \log \left| \det \frac{\partial f_\theta(\mathbf{I})}{\partial \mathbf{I}} \right|, \tag{4.32}$$

where the absolute determinant of the Jacobian appears as a result of the change-of-variables formula. Flow blocks are designed such that the Jacobian are easy to compute since affine coupling

layers make the Jacobian triangular which reduces the determinant to the product of diagonal terms (Dinh et al., 2017).

Flow based models are trained by maximizing this likelihood over the training data, thereby learning a transformation that maps in-distribution data to a Gaussian distribution in latent space. NFs differ from VAEs in a critical way. While both map data into a latent space, VAEs typically rely on an approximate objective for the likelihood, whereas flows can compute the log-likelihood exactly. The reason for this is that VAEs use a non-invertible encoder and therefore only provide a proxy for the true likelihood, while flows are invertible and therefore preserve all information.

Each flow model is composed of multiple invertible blocks f_k, which can be understood as special neural network layers designed to be reversible transformations. A common implementation of this are the affine coupling layers (Dinh et al., 2017). In affine coupling layers, the feature map is split into two parts. One part remains unchanged while the other neural network uses it to compute how to scale and shift the other part.

Due to the fact that this transformation is reversible, many of these can be stacked together. If we compose L such blocks the full transformation is

$$\mathbf{z} = f_L \circ f_{L-1} \circ \cdots \circ f_1(\mathbf{I}), \tag{4.33}$$

meaning that the input image is passed sequentially through all invertible layers.

Anomaly detection models are trained on in-distribution images. During training the flow learns to assign high likelihood to in-distribution examples. At inference, samples that differ from this learned normality provide low likelihood values, which can be used as anomaly scores:

$$s(I) = -\log p_I(I), \tag{4.34}$$

where $s(I)$ is the anomaly score derived from the models estimated likelihood.

While Eq. (4.34) defines an slice-level anomaly score, flow-based models such as FastFlow and CFlow also yield pixel-level information. Their feature extractors produce multi-scale feature maps, and a normalizing flow is applied to each spatial location of these maps, generating a likelihood per pixel.

4.5.1 FastFlow

FastFlow builds upon the foundation of normalizing flows by extending them into two spatial dimensions, thereby addressing key challenges of earlier flow-based anomaly detection approaches such as DifferNet (Rudolph et al., 2021). In previous works, the two-dimensional feature maps extracted from convolutional backbones were flattened into one-dimensional vectors before being processed by the flow model. This early design motivated the development of FastFlow, as its operation disrupts spatial information, making it difficult to capture local structures and to produce meaningful pixel-level anomaly maps. FastFlow solves this issue by creating a dedicated 2D normalizing flow design. This preserves the spatial layout of features and enables end-to-end anomaly localization (Yu et al., 2022).

Figure 4.9 summarizes the overall FastFlow pipeline and the structure of one flow step. FastFlow first employs a frozen feature extractor $\phi(\cdot)$ usually being a ResNet or Vision Transformer in order to obtain $\mathbf{F} = \phi(\mathbf{I})$ from the input image $\mathbf{I}$. These feature maps are then passed through a layer of normalizing flows $f_\theta : \mathbf{F} \to \mathbf{Z}$, which transforms the feature maps into a latent into a standard Gaussian in latent space $\mathcal{N}(0, I)$. Each block f_k is composed of multiple affine coupling layers. The coupling operation itself is invertible while small convolutional subnetworks inside the block only predict scale and shift parameters used by the affine transformation. This ensures that the overall mapping is invertible even if the subnetworks themselves are not. By alternating 3×3 and 1×1 convolution kernels, the model can capture both fine-grained local context

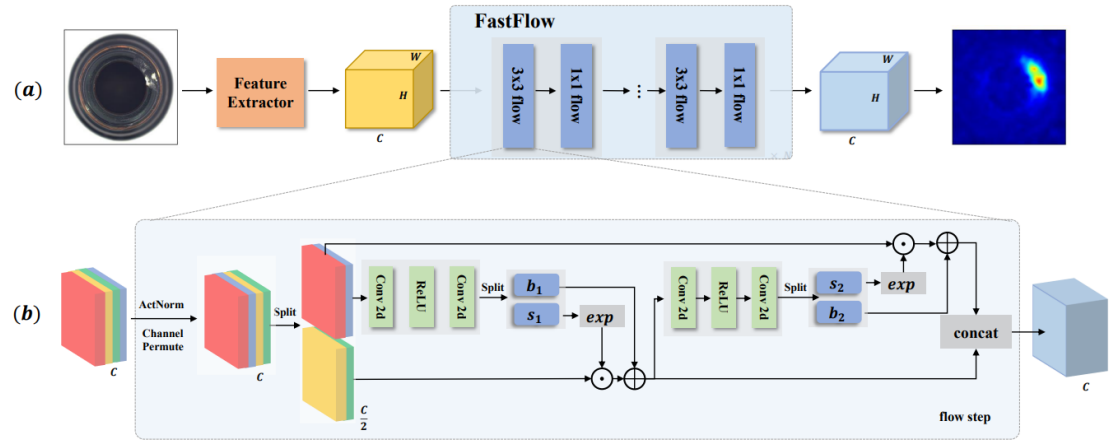


Figure 4.9: FASTFLOW ARCHITECTURE. *Overview of the FastFlow architecture, taken from Yu et al. (2022). Panel (a) shows the overall pipeline: a frozen CNN backbone extracts a feature map which is then processed by 2D normalizing-flow blocks to produce an anomaly likelihood map. Panel (b) illustrates a singular flow step where the feature channels are normalized, split, and transformed via affine coupling before being recombined into an updated C-channel tensor.*

and broader global relationships. This structure captures spatial information while allowing the computation of pixel-level likelihoods directly, without flattening (Yu et al., 2022).

During training, FastFlow is optimized by minimizing the negative log-likelihood of the backbone feature maps extracted from in-distribution images. Specifically for each training image $\mathbf{I}$ with feature map $\mathbf{F} = \phi(\mathbf{I})$, the loss is computed as introduced in Eq. (4.32).

At test time, FastFlow applies the likelihood computation from Eq. (4.32) not only at an slice-level but rather independently to each C-dimensional feature vector $\mathbf{F}_{h,w}$ in the downsampled feature map. This produces a grid of negative log-likelihood values $S_{h,w}$, which serves as the pixel-wise anomaly score map. Regions with high scores are considered likely to contain anomalies. Because the feature map is spatially smaller than the input, image S is upsampled using bilinear interpolation thereby providing detailed anomaly localization (Yu et al., 2022).

FastFlow achieves a favorable balance between speed and accuracy through its compact architecture and single-pass inference pipeline. Unlike earlier flow-based methods that relied on patch-wise evaluation or sliding-window schemes, FastFlow processes the entire feature map at once, which substantially reduces inference time while retaining strong localization capability (Yu et al., 2022; Gudovskiy et al., 2022). FastFlow´s efficiency and modular design make it suitable for a flow-based model.

4.5.2 CFlow

CFlow (Gudovskiy et al., 2022) is a flow-based model designed for unsupervised anomaly detection and localization.Figure 4.10 extends the flow framework by introducing conditional flows. These allow to incorporate spatial priors during likelihood estimation.

As in FastFlow, CFlow is built from a frozen convolutional backbone and a trainable flow head. The backbone $\phi(\cdot)$ extracts multi-scale feature maps $\mathbf{F}_k = \phi_k(\mathbf{I})$ from the input image $\mathbf{I}$. These feature maps capture local patterns as well as larger structural elements in the in-distribution MR slices. What distinguishes CFlow is that each feature level k is processed by a conditional

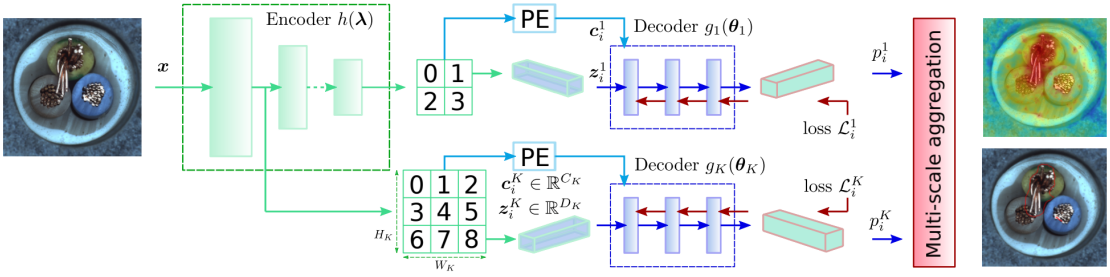


Figure 4.10: CFLOW ARCHITECTURE. *Overview of the CFlow architecture, taken from Gudovskiy et al. (2022). An input image* $\mathbf{I}$ *is encoded into multi-scale feature maps; at each scale* k*, flattened feature vectors* z_i^k *are combined with positional codes* c_i^k *and passed through a conditional flow decoder* $g_k(\theta_k)$ *trained via negative log-likelihood. At inference, per-location anomaly scores from all scales are aggregated into a final heatmap.*

normalizing flow f_{θ_k}, which models the density of the normal features using a spatial context variable $c_k[h, w]$. This conditioning lets the flow take the features location into account, ensuring that the learned distribution reflects how normal features typically show up in different positions of an image.

During inference, CFlow applies its conditional flow f_{θ_k} to every feature vector $\mathbf{F}_k[h, w]$ in the feature map. For each spatial location (h, w), the flow transforms the feature into a latent vector and evaluates its log-likelihood under the Normal distribution. The negative log-likelihood becomes the pixel-wise anomaly score.

Because this operation is performed independently at each feature level k, CFlow produces one anomaly map $\mathbf{M}_k$ per scale. Each of these maps are then upsampled to the input size using bilinear interpolation. The final anomaly map is acquired by aggregating the multi-scale maps by element-wise summation:

$$\mathbf{S}_{\text{CFlow}} = \sum_{k=1}^{K} \Psi\big(\mathbf{M}_k\big), \tag{4.35}$$

where $\Psi(\cdot)$ is the bilinear upsampling to the input size, and K is the number of feature scales (Gudovskiy et al., 2022). Locations that are unlikely under the learned normal distribution at any scale obtain high scores in $\mathbf{S}_{\text{CFlow}}$.

Chapter 5

Materials and Methods

In this chapter, we describe the data used to develop and evaluate our MR-only quality assurance pipeline. We first summarize the open-source SynthRAD2023 and SynthRAD2025 datasets, which provide paired MR–CT studies of brain and pelvis regions. We then detail our annotation strategy for defining OOD slices directly on MR, supported by CT for context, and outline the guidelined followed during the annotation process. Finally, we explain how the raw data were prepared for our experiments, and clarify the scope we adopted for training and evaluation, with a focus on the pelvis region and implant-related anomalies. This structure is intended to clearly link dataset design choices to the experimental questions addressed in the following sections.

5.1 Dataset

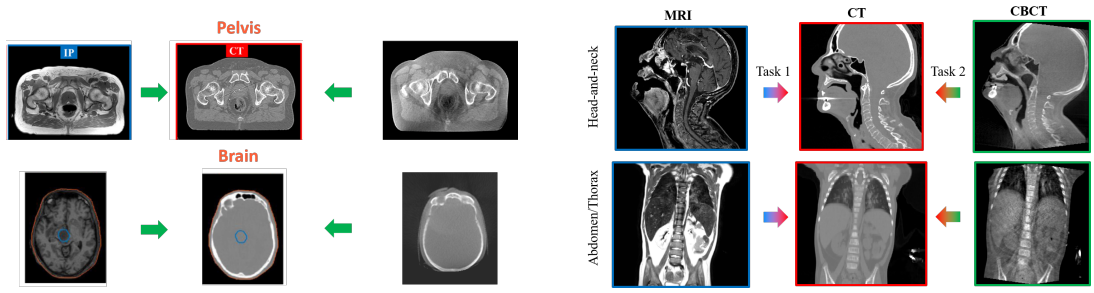
5.1.1 Overview of SynthRAD2023 and SynthRAD2025 Datasets

In this study, we use the MR-to-CT datasets from SynthRAD2023[1] and SynthRAD2025[2] Grand Challenges which provide large-scale, multi-institutional data specifically curated for evaluating sCT generation models.

The SynthRAD2023 dataset includes 1,080 cases (brain and pelvis), consisting of MR, CBCT, and CT scans collected from three Dutch academic hospitals (UMC Utrecht, UMC Groningen, and Radboud Nijmegen). Its successor, SynthRAD2025, expands anatomical diversity by incorporating head-and-neck, thorax, and abdominal regions, with a total of 2,362 cases across five European centers (UMC Groningen, UMC Utrecht, Radboud UMC in Netherlands, LMU University Hospital Munich, and University Hospital of Cologne in Germany). Both datasets have undergone preprocessing steps, including image registration, defacing (for brain scans), resampling, and quality control (Thummerer et al., 2023, 2025). Data are provided in standardized formats: NIfTI for 2023 and MetaImage for 2025, accompanied by body contour masks and acquisition metadata. Training and validation sets are publicly accessible during the challenge period; test sets remain embargoed to ensure fair benchmarking. Table 5.1 summarizes the key characteristics of the data set.

These datasets are released with the goal of fostering reproducible, transparent evaluation. For our current work, both datasets are specifically designed to support the development and evaluation of MR-to-CT image synthesis models in clinically relevant scenarios. By providing paired MR and CT scans across multiple institutions, anatomical sites, and acquisition protocols, they offer a valuable foundation to study model robustness and generalization to OOD inputs.

[1] https://synthrad2023.grand-challenge.org/
[2] https://synthrad2025.grand-challenge.org/synthrad2025/



(a) SynRAD2023 (b) SynRAD2025

Figure 5.1: REPRESENTATIVE EXAMPLES FROM SYNTHRAD2023 AND SYNTHRAD2025 DATASETS. *Subfigure (a) and (b) show tasks from the challenges, respectively. Each subfigure includes examples of both Task 1: MR-to-sCT synthesis to support MR-only and MR-based adaptive (only for SynRAD2025) radiotherapy, and Task 2: CBCT-to-sCT synthesis to facilitate CBCT-based adaptive radiotherapy (Thummerer et al., 2023, 2025).*

Table 5.1: DATASET SUMMARY. *This table provides an overview of datasets SynthRAD 2023 and 2025 Challenges' characteristics relevant to this study.*

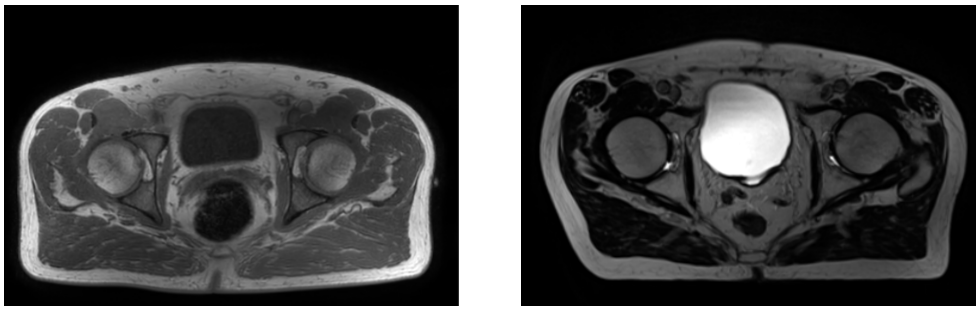
Feature	Datasets	
	SynRAD2023	**SynRAD2025**
Total cases	1,080	2.362
Modality pairs (MR–CT/CBCT–CT)	540 / 540	890 / 1,472
Anatomical coverage	Brain, pelvis	HN, thorax, abdomen
Age range	3-93 years	Not reported
Data format	NIfTI (.nii.gz)	MetaImage (.mha)
Image registration	Rigid only	Rigid, deformable
Split (Trn/val/test approx. in %)	67 /11 / 22	65 / 10 / 25
CT ground truth embargo	Test set hidden	Released by 2030

The pelvis SynthRAD2023 dataset consists of scans from two different centers A and C, which employed two two different techniques to acquire the image data. Center A utilized a T1-weighted spoiled gradient echo (FFE) sequence, and center C used a T2-weighted fast spin echo (SPACE) sequence (Thummerer et al., 2023). A T1-weighted image looks different from a T2-weighted image. T1 highlights fat and subacute hemorrhage as bright, while water appears dark. T2 highlights fluid/water (edema, cysts) as bright. The difference in appearance of these different techniques can be seen in Figure 5.2, which displays pelvic MR images from Center A (left) and Center C (right), both from the SynthRAD2023 dataset. This figure exemplifies how scanner and protocol variability leads to differences in visual representation of the same anatomy.

5.1.2 Dataset Annotation

OOD Classes

Based on organized discussions with the Medical Physics group at the University Hospital of Zurich, we identify a focused set of OOD classes. These classes capture patterns that are likely to degrade sCT generation. The emphasis during annotation is kept on what is visible on MR slices



(a) Center A (b) Center C

Figure 5.2: COMPARISON OF PELVIC MR IMAGES. *Axial pelvic MR images from the SynthRAD2023 multi-center dataset. The subfigure (a) was acquired at Center A, and (b) at Center C. The scans illustrate how MR appearance can vary depending on acquisition site and protocol.*

in terms of the effects of artifacts or anomalies, since the AD models are trained purely on MR data.

Implant-Metal Metallic components (e.g., orthopedic implants, surgical clips, screws, brachytherapy seeds) produces characteristic dark signal voids on MR and strong scatter on CT, often obscuring nearby anatomy. Because their impact is consistent and substantial, any slice showing a metallic implant is labeled as OOD. An example of this OOD category is illustrated in Figure 5.3.

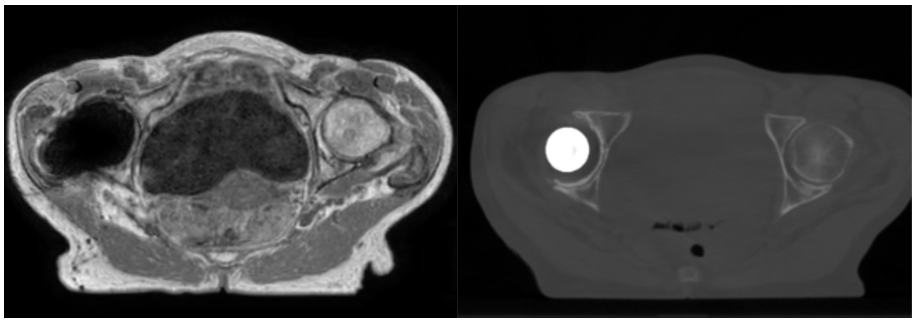


Figure 5.3: VISUAL EXAMPLE OF THE IMPLANT-METAL CATEGORY. *The MR scan (left) displays the characteristic artifacts defining the Implant-Metal category, while the paired CT scan (right) serves as a reference.*

Bright Hotspot in MR Slices with unexpected focal bright signals on MR that are not characteristic of in-distribution anatomy are labeled OOD. These hotspots commonly reflect artifacts or unusual signal behavior. An example of this OOD category is illustrated in Figure 5.4.

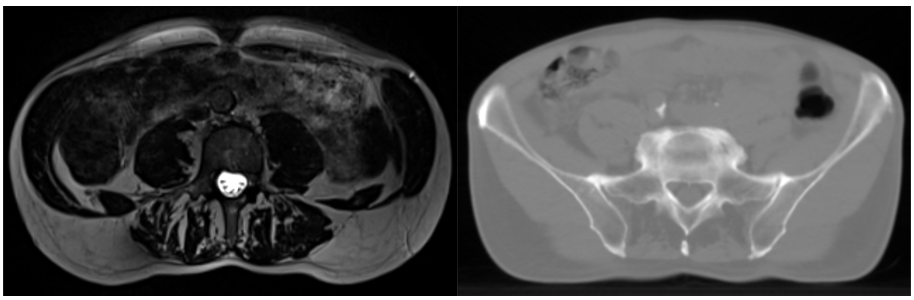


Figure 5.4: VISUAL EXAMPLE OF THE BRIGHT HOTSPOT IN MR CATEGORY. *The MR scan (left) displays the characteristic artifacts defining the Bright Hotspot in MR category, while the paired CT scan (right) serves as a reference.*

Contrast Agent Sequences exhibiting high-signal patterns in the vasculature and surrounding tissues, indicative of contrast uptake, are grouped here when this brightening is distinctly visible on both MR (e.g., gadolinium-enhanced T1) and paired CT. If enhancement is evident only on CT but not observed on MR, the case is not labeled OOD since the learning is done purely on MR data. An example of this OOD category is illustrated in Figure 5.5.

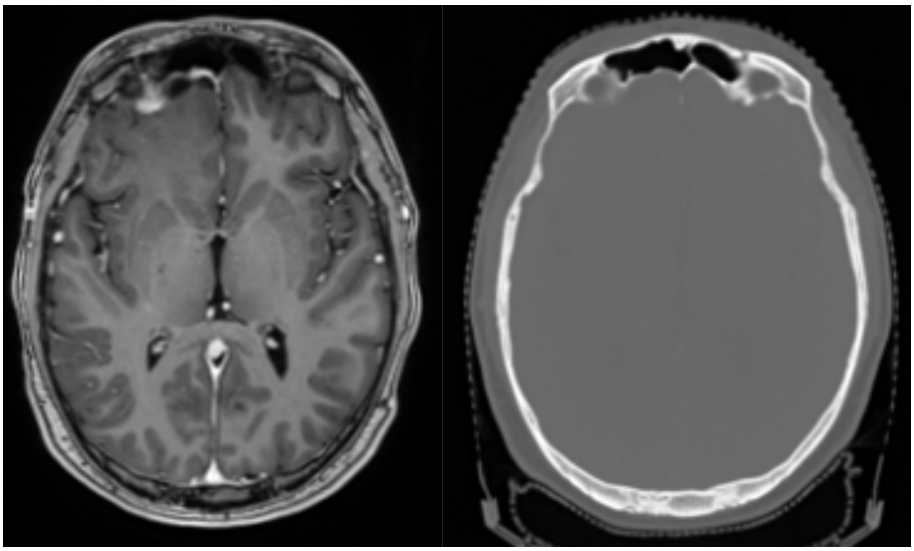


Figure 5.5: VISUAL EXAMPLE OF THE CONTRAST AGENT CATEGORY. *The MR scan (left) displays the characteristic artifacts defining the Contrast Agent category, while the paired CT scan (right) serves as a reference.*

Unseen Anatomy-Missing Bone Post-surgical absence of bony structures (e.g., bone flap not replaced) is labeled OOD when it creates marked asymmetry or disrupted continuity. If metal clips are present around fracture lines, those slices are included in the *Implant-Metal* category. An example of this OOD category is illustrated in Figure 5.6.

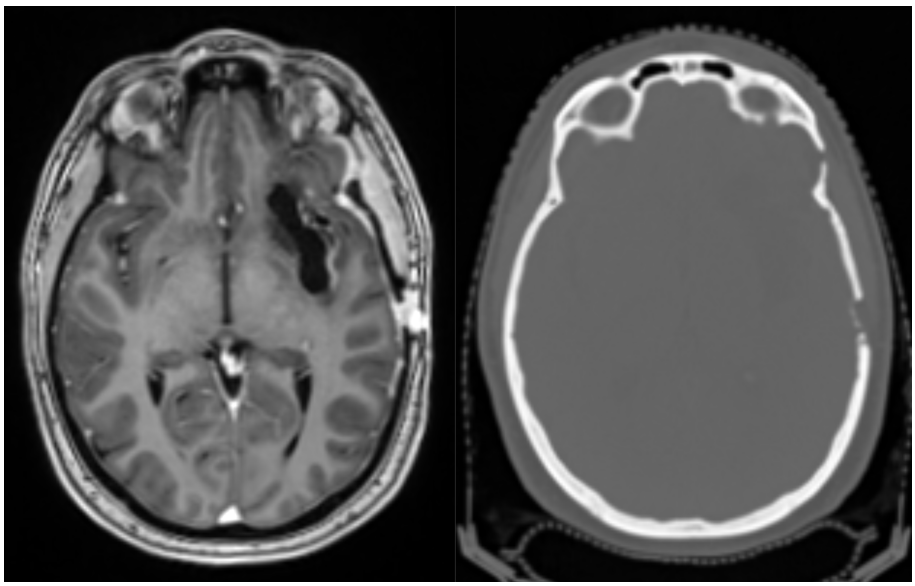


Figure 5.6: VISUAL EXAMPLE OF THE UNSEEN ANATOMY-MISSING BONE CATEGORY. *The MR scan (left) displays the characteristic artifacts defining the Unseen Anatomy-Missing Bone category, while the paired CT scan (right) serves as a reference.*

MR Markers External liquid-based positioning markers appear as bright signals on MR and may persist after preprocessing. Because they can introduce high-intensity regions that mislead sCT generation, such sequences are labeled OOD. An example of this OOD category is illustrated in Figure 5.7.

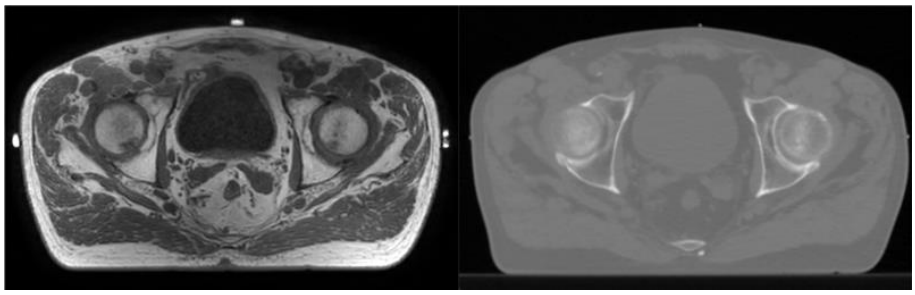


Figure 5.7: VISUAL EXAMPLE OF THE MR MARKERS CATEGORY. *The MR scan (left) displays the characteristic artifacts defining the MR Markers category, while the paired CT scan (right) serves as a reference.*

Other A restricted, catch-all class for rare or ambiguous cases (e.g., unusual anatomical variants or atypical artifacts) that do not fit any of the the OOD categories. Consequently, these cases are removed from the experimental datasets to preserve labeling consistency and ensure the quality of the annotation process.

Annotation Rules and Results

Our procedure follows the MR-only detection objective, which focuses on the labels that reflect OOD patterns that are actually visible on MR. The corresponding CT scans are consulted to support interpretation and validation, but visibility on MR slice is decisive factor for OOD labeling of the slices.

We adhere to an **MR-first criterion** where a case is labeled OOD only if the anomaly is detectable on MR slices. Findings seen exclusively on CT are not labeled OOD since the training of the AD models relies purely on MR data. Serving as supportive context, CT helps to the validation of MR findings (e.g., metal scatter patterns, vascular enhancement) without overriding the MR-visibility rule. Regarding **registration tolerance**, MR and CT are not perfectly aligned; small slice offsets are expected. Annotators account for typical misalignment when cross-checking across modalities. For **reference view and indexing**, annotations are defined in the axial plane to ensure consistent slice-range reporting across studies and annotators. Finally, allowing for **multiple findings per 3D volume**, distinct OOD occurences within within the same 3D volume are recorded independently, so that each OOD event with its own class and slice range can be analyzed in isolation.

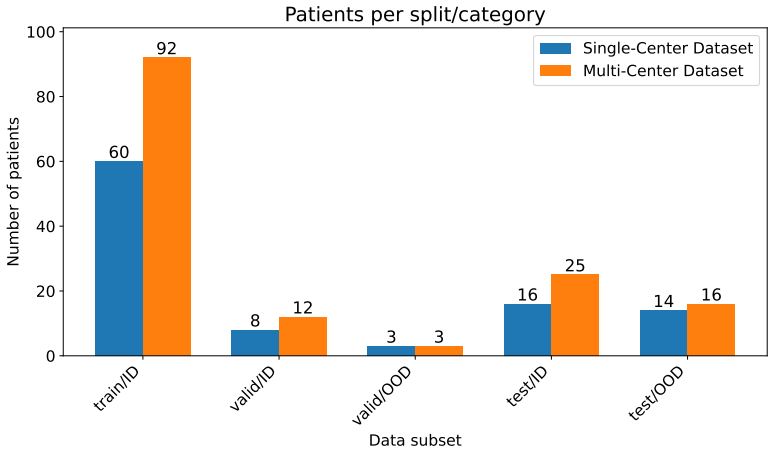
5.1.3 Final Experimental Datasets

After annotation of all relevant out-of-distribution classes, we refined the experimental scope to include only pelvic MR scans from the SynthRAD2023 dataset. In consultation with clinical experts, the OOD category for the core study is limited to a single, well-defined group: *Implant-Metal*. Accordingly, two distinct classes are represented in each dataset: in-distribution pelvis MR studies and OOD cases with metallic implants in the pelvis. Two datasets are constructed: a single-center dataset and a multi-center dataset. In both, images are split into training, validation, and test sets at the patient-level, ensuring that no patient's scans appear in more than one set. Only in-distribution scans are included in training, while validation and test sets include both in-distribution and OOD scans. For the pixel- and slice-level metrics, these test sets are used.

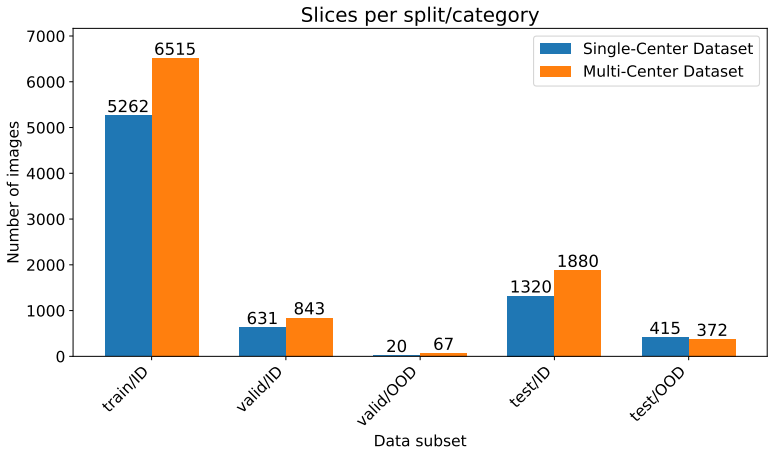
For pixel- and slice-level metrics, we evaluate models on the test sets defined above, using only the slices that fall within the annotated OOD ranges for OOD volumes. For patient-level metrics, we instead use a setting that reflects clinical use, in which all slices of in-distribution volumnes are included, and for OOD cases, the entire MR volume is submitted (not only slices with visible artifacts). In practice, an automated system must decide whether a volume is OOD based on the full scan.

Single-Center Dataset The primary dataset consists of pelvic MR scans from center A only. This single-center dataset is used to ensure cohort homogeneity and to minimize confounding from protocol variability in the primary experiments. This single-center dataset is composed of 101 patient scans and $7'648$ individual slices. Out of all scans, 17 are marked as OOD. The distribution of the single-center dataset is shown in Figure 5.8.

Multi-Center Dataset For a broader evaluation, the multi-center dataset includes pelvic MR scans from centers A and C of the SynthRAD2023 dataset. The addition of center C introduces increased heterogeneity and enables analysis of performance in more varied clinical settings. This dataset is composed of 146 patient scans and $9'679$ slices. Out of these scans, 19 are marked as OOD, meaning 2 new OOD volumes compared to the single-center dataset. The distribution of the multi-center dataset can be seen in Figure 5.8.



(a) Patients per data subset



(b) Images per data subset

Figure 5.8: DISTRIBUTION OF THE SINGLE-CENTER AND MULTI-CENTER DATASETS. *The bar chart in (a) shows the number of patient scans per data subset (train/validation/test) for both the single-center and multi-center datasets, and (b) shows the corresponding number of image slices per data subset.*

5.2 Ground-Truth Anomaly Mask Generation

It is challenging to directly extract metal anomaly masks from MR images, because metal-induced susceptibility artifacts cause signal voids and pile-ups that extend beyond the true implant boundaries. However, on CT images, metal appears with consistently very high Hounsfield Unit values, providing a signal that allows segmentation using simple threshold strategies. Therefore, we adopted a pipeline to first extract the candidate anomaly masks from pre-registered CT scans, which were further refined and extended to full artifact region within the MR domain. This final anomaly mask served as the ground truth for labeling anomalous regions in our experiments.

5.2.1 Intensity-based candidate CT mask

Dense materials like bone and metal exhibits high X-ray attenuation, and appears bright on the CT image. A straightforward approach to mark anomalous regions with metal is to use a fixed threshold, (e.g., HU > 2000). However, scans from SynthRAD2023 are acquired from multiple centers with heterogeneous scanning protocols and reconstruction parameters. Furthermore, the optimal threshold can vary between scans from the same center in practice. A single global threshold fails to generalize. To overcome the limitations of a fixed HU threshold, we implemented an adaptive technique that computes a patient-specific threshold by analyzing the maximum HU values in pre-annotated in-distribution and OOD slices, to ensure an optimal threshold for each scan.

The following steps are applied to get the candidate metal mask for each CT scan.

1. **Slice-wise HU Profile Analysis**: The 3D CT volume is analyzed along the axial plane to construct an intensity profile. For each slice z, the maximum HU value is computed.

$$\mathtt{HU}_{\text{max}}(z) = \max_{x,y} \mathtt{HU}(x, y, z) \tag{5.1}$$

2. **Adaptive Scan Threshold Estimation**: A scan threshold τ_{scan} is derived to separate in-distribution from OOD slices. Let $\mathcal{N}$ be the set of in-distribution slices and $\mathcal{A}$ the set of OOD slices. The threshold is defined as:

$$\tau_{\text{scan}} = \max\left(\max_{z\in\mathcal{N}} \mathtt{HU}_{\text{max}}(z), \min_{z\in\mathcal{A}} \mathtt{HU}_{\text{max}}(z)\right) \tag{5.2}$$

3. **Sensitivity Adjustment**: A sensitive buffer Δ is applied to account for potential intensity underestimation. The buffer value is determined based on experience as 250.

$$\tau_{\text{final}} = \tau_{\text{scan}} - \Delta \tag{5.3}$$

4. **Binary Mask Generation**: A candidate CT mask $M_{\text{candidate}}$ is generated by thresholding the entire CT volume at τ_{final}:

$$M_{\text{candidate}}(x, y, z) = \begin{cases} 1 & \text{if } \mathtt{HU}(x, y, z) \geq \tau_{\text{final}} \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

5.2.2 MR-Guided Flood Fill for Artifact Extension

The initial candidate mask $M_{\text{candidate}}$, derived from the CT thresholding, identifies the high-density metal object. However, the true artifact region in the co-registered MR volume often extends beyond the object itself, appearing as signal voids. To extend $M_{\text{candidate}}$ to encompass this full artifact

area visible in the MR image, we implemented an MR-Guided Flood Fill technique, leveraging the distinct intensity changes of the artifact boundary in the MR signal. This refinement procedure was applied slice-wise:

1. **Contour and Centroid Detection**: The external contours were extracted from the binary CT mask $M_{\text{candidate}}$ on the current slice. For each contour, the centroid (c_x, c_y) was calculated, serving as the seed point for the flood fill.

2. **Small Area Exclusion**: To prevent small, likely non-metal false positive regions from introducing noise, any contour with an area below a set threshold ($\texttt{min\_contour\_area} = 5$ pixels) was preserved directly from $M_{\text{candidate}}$ and excluded from the flood-fill process.

3. **Adaptive Flood Fill**: The `cv2.floodFill` algorithm was applied to the current MR image slice. Let $I(x, y)$ denote the image intensity at pixel location (x, y) on this slice. Starting from the centroid (c_x, c_y), connected pixels were included in the anomaly mask if their intensity lay within a specified range around the seed intensity $I(c_x, c_y)$. The sensitivity was controlled by a lower bound difference ($\texttt{lo\_diff} = 5$) and an upper bound difference ($\texttt{up\_diff} = 10$):

$$I(c_x, c_y) - \texttt{lo\_diff} \leq I(x, y) \leq I(c_x, c_y) + \texttt{up\_diff} \tag{5.5}$$

4. **Mask Aggregation**: All binary masks generated by the successful flood-fill operations were combined with, if present, the small, preserved segments of $M_{\text{candidate}}$ to form a preliminary refined mask for the slice.

This process effectively expanded the mask based on the spatial and intensity characteristics of the MR artifact, providing an initial MR-refined mask M_{refined}.

5.2.3 Conditional Morphological Mask Processing

Morphological operations provide basic shape-manipulation tools for binary images (Soille, 2003). Let $M \subset \mathbb{Z}^2$ denote the foreground pixel set of a binary mask and $B \subset \mathbb{Z}^2$ a finite set of offsets, called the structuring element (for example, a small disk- or square-shaped neighborhood). *Dilation* of M by B, written $M \oplus B$, expands foreground regions by marking a pixel z as foreground if any translated copy of B centered at z overlaps M, thereby enlarging objects and bridging small gaps. *Erosion*, denoted $M \ominus B$, shrinks foreground regions by keeping a pixel z foreground only if the translated structuring element fits entirely inside M, which removes thin protrusions and can separate narrowly connected objects. By composing these primitives, *opening* and *closing* are defined as

$$M \circ B = (M \ominus B) \oplus B, \qquad M \bullet B = (M \oplus B) \ominus B. \tag{5.6}$$

Opening removes small isolated foreground regions and smooths object boundaries, whereas closing fills small holes and connects nearby components (Soille, 2003). In practice, these operations are implemented by sliding a binary kernel B across the mask, as in standard image-processing libraries.

The MR-refined mask M_{refined} still contains minor boundary irregularities and small spurious regions, necessitating a final cleaning step. Morphological opening is well-suited for smoothing contours and removing small isolated objects when applied with an appropriate structuring element B. However, applying opening uniformly across all slices risks undesirably eroding or even removing small, genuine metal artifact segments. Conditioning the operation on the size of individual connected components was also found to be suboptimal, because large artifacts were insufficiently smoothed, while small, noisy regions in their vicinity were retained.

To balance preservation and smoothing, we adopt a conditional, slice-wise opening based on the total artifact area in the current slice. For each slice of M_{refined}, we first compute the total number of foreground pixels (artifact area). A morphological opening with a circular structuring element B (implemented as a binary disk-shaped kernel with diameter $\texttt{disk\_size} = 5$ pixels) is then applied only if this area exceeds a minimum threshold $\texttt{min\_area\_for\_smooth} = 50$ pixels. Formally, the conditional opening is defined as:

$$M_{\text{final}} = \begin{cases} \texttt{Open}(M_{\text{refined}}; B) & \text{if } \texttt{Area}(M_{\text{refined}}) \geq 50 \\ M_{\text{refined}} & \text{otherwise} \end{cases} \tag{5.7}$$

Here, $\texttt{Open}(...; B)$ denotes an opening with the disk-shaped structuring element B corresponding to the circular kernel used in the implementation. This conditional strategy smooths the boundaries of large, high-confidence artifacts while preserving the structure of small metal artifact segments, yielding the final artifact mask M_{final}.

The results of the mask refinement procedure are visualized in Figure 5.9. For each patient, the MR slice, raw CT-based mask, and refined mask are shown side by side to facilitate qualitative assessment of the improvement of the ground truth anomaly mask. The figure illustrates that accurate ground truth anomaly masks can be generated through the refinement process. The remaining small discrepancies are primarily attributable to limitations in the initial CT-based segmentation, which cannot be completely resolved by the refinement alone.

5.3 MR Image Preprocessing

All MR images are first processed to exclude background and non-body regions. For each scan, a coarse body mask is generated on the MR magnitude volume by thresholding low-intensity background values in the MR image and refining the result with binary erosion, largest-component selection, and binary dilation. This ensures that only anatomically relevant areas are retained for further analysis.

The body mask is applied slice-wise to the MR volume, and the masked MR is then min–max normalized to the range $[0, 255]$. Original in-plane resolutions vary slightly across scans, so a fixed input size is required for all models. To standardize inputs while keeping the body fully in view, each masked slice is first center-padded to a square, resized to 240×240 pixels, and then center-cropped to a final size of 224×224. The final size of 224×224 matches the typical input resolution of ImageNet-pretrained backbones, which most models used in our experiments. The padding–resize–crop sequence removes residual background while harmonizing the field of view across patients.

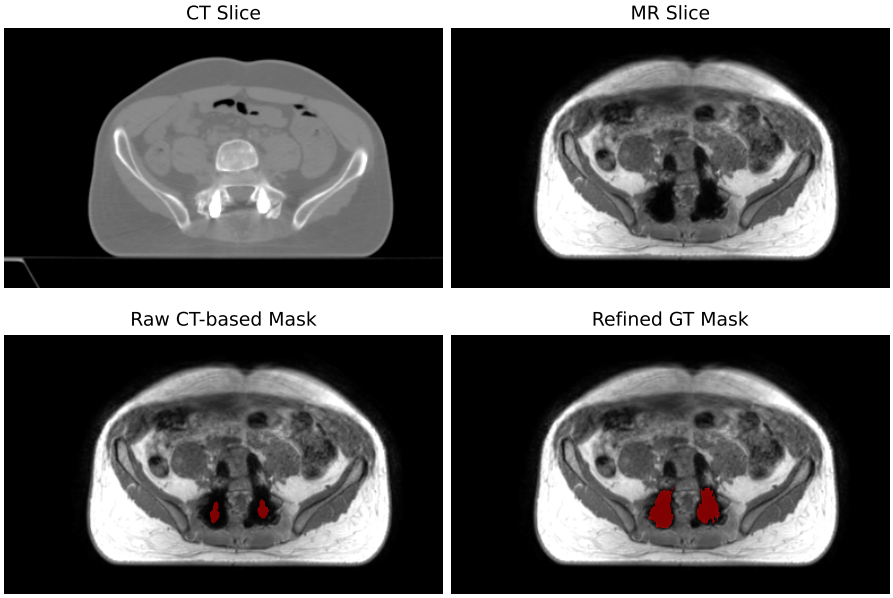
Based on these preprocessed slices, three dataset variants are generated to support different model architectures and input conventions.

Replicated-Channel Slices

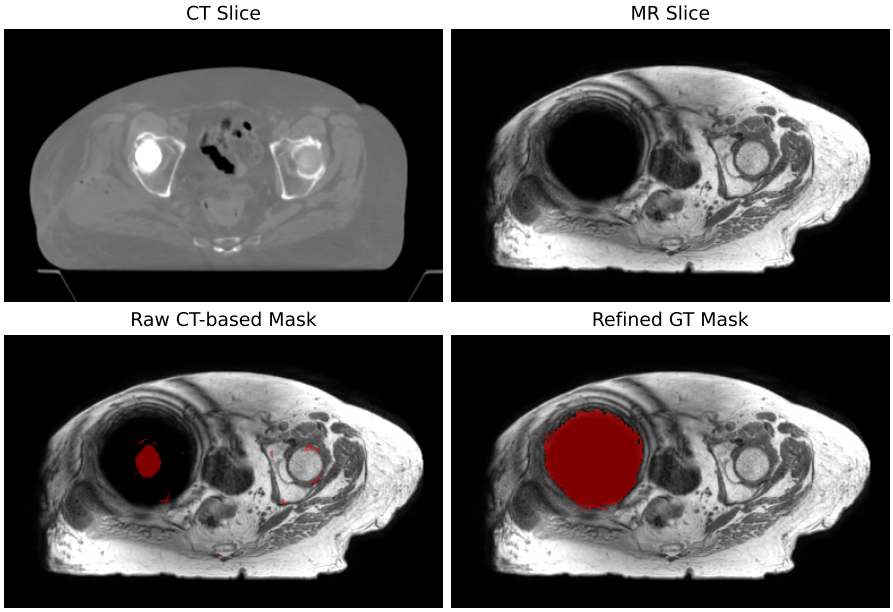
In the first variant, each preprocessed MR slice is replicated across three channels and stored as a three-channel NIfTI volume. This representation preserves voxel-level fidelity in the NIfTI format while matching the three-channel input convention of many convolutional architectures. All three channels contain identical intensity information.

Colormap-Encoded Slices

In the second variant, each preprocessed MR slice is stored as a three-channel RGB image. The normalized slice is replicated across three channels and saved using the `bone` colormap. The

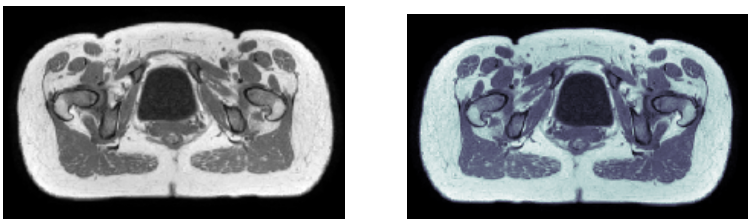


(a) Aanomaly mask refinement for volume 1PA136 slice 110.



(b) Anomaly mask refinement for volume 1PA155 slice 83.

Figure 5.9: COMPARISON OF CT/MR SLICES WITH RAW AND REFINED ANOMALY MASKS. *Side-by-side comparison of corresponding CT and MR slices with raw and refined artifact masks for two example patients. From left to right, top to bottom: (i) the CT slice, (ii) the MR slice, (iii) the MR slice with the raw CT-based mask overlaid, and (iv) the MR slice with the refined ground-truth mask overlaid. Mask regions are shown in red to highlight metal-induced artifacts.*



(a) Replicated Channel Slice (b) Bone Colormap Slice

Figure 5.10: COMPARISON OF INPUT FORMATS. *The subfigure (a) shows the MR slice in replicated channel format, and (b) shows the same MR slice with the bone colormap applied.*

`bone` palette is a grayscale-like colormap commonly used in medical visualization that preserves the relative contrast between tissues while remaining compatible with ImageNet-pretrained CNN backbones that expect three-channel inputs.

Adjacent-Context Slices

In the third variant, three consecutive MR slices (previous, current, next) are stacked along the channel dimension to form a three-channel NIfTI volume. This provides limited adjacent anatomical context, allowing models to exploit short-range inter-slice relationships without requiring full 3D processing.

5.4 Anomalib Model Training

Anomalib is an open-source library that implements state-of-the-art methods for unsupervised anomaly detection and localization with a focus on modular design for training, evaluation, and deployment (Akcay et al., 2022). Most models in this project were trained through Anomalib. This section summarizes (i) model outputs, (ii) how raw anomaly scores are normalized and thresholded by Anomalib's post-processor, and (iii) the model training hyperparameters used in our experiments. The Anomalib GitHub repository is available at [3].

Training hyperparameters To make the experimental setup reproducible, we report the main training hyperparameters and backbone settings for each evaluated method in Table 5.2. Unless stated otherwise, we used the default Anomalib parameters and did not tune hyperparameters per dataset variant. The final training settings were selected by tuning hyperparameters on the validation set.

Model outputs Let a 2D input slice be $I \in \mathbb{R}^{H\times W}$ with a pixel domain of $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$. An Anomalib model produces: **(i) Raw anomaly score map** $S(I) : \Omega \to \mathbb{R}$ (unbounded prior to normalization). For a pixel $u \in \Omega$ we write $S(I,u)$ for its pixel anomaly value. **(ii) Anomaly map** $A(I) : \Omega \to [0,1]$ is obtained by normalizing $S(I)$. **(iii) Prediction mask** $\widehat{M}(I) : \Omega \to \{0,1\}$ is obtained by thresholding the raw anomaly scores at an optimal threshold $\tau^\star$ learned on the validation set.

[3] https://github.com/openvinotoolkit/anomalib

Table 5.2: MODEL-LEVEL TRAINING CONFIGURATION. *Key training hyperparameters and training mode per model. "Decoder/head only" indicates that the backbone is kept fixed and only the model head is optimized. Methods such as PatchCore are non-parametric and therefore do not use gradient-based training.*

Model	Optimizer (trained params)	LR	Batch	Epochs	Notes
RD4AD	Adam (decoder only)	10^{-3}	16	100	backbone fixed
STFPM	SGD	$4 \cdot 10^{-1}$	16	50	teacher fixed
PatchCore	– (non-parametric)	–	32	1	feature extraction + coreset (0.1%), kNN (k=9)
CFA	Adam (default)	10^{-3}	32	100	early stopping: patience = 5 (image AUROC)
FastFlow	Adam (flow head)	10^{-3}	8	250	weight decay = 10^{-5}
CFlow	Adam (decoder only)	10^{-4}	8	50	fiber batch size = 64
DRAEM	Adam	10^{-4}	4	100	synthetic-anomaly training
Dinomaly	StableAdamW	$2 \cdot 10^{-3}$	8	100	frozen DINOv2 encoder
DeepSVDD	Adam	10^{-4}	200	300	AE pretrain: 150 epochs (10^{-5})
CutPaste	SGD	10^{-2}	16	64	3-way variant

Concretely, the post-processor transforms:

$$A(I,u) \;=\; \mathrm{clip}\!\left(\frac{S(I,u)-\tau^\star}{s^{\mathrm{val}}_{\max}-s^{\mathrm{val}}_{\min}}+\tfrac{1}{2},\,0,\,1\right), \qquad \widehat{M}(I,u)=\begin{cases}1, & \text{if } A(I,u)\geq 0.5,\\ 0, & \text{otherwise.}\end{cases} \tag{5.8}$$

where $s^{\mathrm{val}}_{\min}, s^{\mathrm{val}}_{\max}$ are the minimum and maximum raw anomaly scores computed on the validation set. The added $\frac{1}{2}$ recenters the normalized scores such that the learned optimal threshold corresponds to 0.5 in normalized domain. The function $\mathrm{clip}(\cdot,0,1)$ caps values to $[0,1]$ for interpretability and visualization. In Anomalib this normalization and thresholding step is implemented by the post-processor [4].

Learning optimal threshold on validation set Anomalib provides both manual and adaptive thresholding; in our setup we made use of the default adaptive threshold selection. On the validation set $\mathcal{V}$, Anomalib selects an operating threshold $\tau^\star$ by maximizing a criterion J over a set of candidate thresholds:

$$\tau^\star \;\in\; \arg\max_{\tau\in\Theta_{\mathrm{val}}} \; J(\{(S(I),M_{\mathrm{gt}}(I)) : I\in\mathcal{V}\},\,\tau)\,, \tag{5.9}$$

where Θ_{val} denotes the candidate thresholds induced by the raw validation scores. In our case, we use $J=F_1$ computed from pixel-wise counts and defined in Equation 5.15.

Let $M_{\mathrm{gt}}(I,u)$ denote the ground-truth mask for slice I. For slice-level metrics we determine a binary label $y^{\mathrm{img}}(I)\in\{0,1\}$ by setting $y^{\mathrm{img}}(I)=1$ if $\sum_{u\in\Omega} M_{\mathrm{gt}}(I,u)>0$, and 0 otherwise.

Metrics used for threshold selection During threshold selection we define $\widehat{M}_\tau(I,u)$ as a binary prediction mask obtained by thresholding the raw score $S(I,u)$ at a threshold $\tau\in\Theta_{\mathrm{val}}$. After selecting $\tau^\star$, we write $\widehat{M}(I,u)\equiv\widehat{M}_{\tau^\star}(I,u)$.

Across all pixels $u\in\Omega$ of all validation images $I\in\mathcal{V}$, we define:

$$\mathrm{TP}(\tau) \;=\; \Big|\big\{(I,u)\in\mathcal{V}\times\Omega \;:\; \widehat{M}_\tau(I,u)=1 \text{ and } M_{\mathrm{gt}}(I,u)=1\big\}\Big|, \tag{5.10}$$

$$\mathrm{FP}(\tau) \;=\; \Big|\big\{(I,u)\in\mathcal{V}\times\Omega \;:\; \widehat{M}_\tau(I,u)=1 \text{ and } M_{\mathrm{gt}}(I,u)=0\big\}\Big|, \tag{5.11}$$

$$\mathrm{FN}(\tau) \;=\; \Big|\big\{(I,u)\in\mathcal{V}\times\Omega \;:\; \widehat{M}_\tau(I,u)=0 \text{ and } M_{\mathrm{gt}}(I,u)=1\big\}\Big|. \tag{5.12}$$

[4] https://anomalib.readthedocs.io/en/v2.0.0/markdown/guides/how_to/models/post_processor.html

Precision, Recall, and F_1.

$$\text{Precision}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FP}(\tau)}, \tag{5.13}$$

$$\text{Recall}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)}, \tag{5.14}$$

$$F_1(\tau) = \frac{2\,\text{Precision}(\tau)\,\text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} = \frac{2\,\text{TP}(\tau)}{2\,\text{TP}(\tau) + \text{FP}(\tau) + \text{FN}(\tau)}. \tag{5.15}$$

If a denominator is zero, the corresponding metric is 0.

Classification Let $A(\mathbf{I}_{p,z}, u)$ be the normalized anomaly-map value at pixel $u \in \Omega$ for slice z of patient p, and let $\widehat{M}_{\text{pp}}(\mathbf{I}_{p,z}, u) \in \{0, 1\}$ be the corresponding post-processed binary prediction mask. We define a slice as anomalous if $\widehat{M}_{\text{pp}}(\mathbf{I}_{p,z}, u) = 1$ for at least one pixel $u \in \Omega$. Patient-level decisions are then made by aggregating the slice-wise outputs as defined in Section 5.6.3.

5.5 Post-Processing

During inference, the model outputs a continuous anomaly score map for each pixel. A binary prediction mask is subsequently generated via thresholding for evaluation purposes. These two forms of output follow distinct downstream processing steps and are used by different evaluation metrics (see Figure 5.11).

Anomaly map branch. Continuous anomaly scores are masked by the body region and are not post-processed. They are used to generate the binary prediction masks that form the basis for all reported evaluation metrics.

Prediction mask branch. Binary prediction masks are post-processed to improve spatial coherence and clinical interpretability. The pipeline removes small isolated regions, smooths boundaries through morphological operations, and enforces cross-slice consistency to ensure that detected anomalies persist across neighboring slices. This yields patient-wise volumes suitable for downstream evaluation with threshold-dependent metrics. All models produce slice-based masks by default, which allows uniform post-processing regardless of input modality (2D slices, 2D NIfTI, or 2.5D NIfTI volumes).

5.5.1 Slice-wise post-processing

We first refine the prediction masks slice-by-slice using purely 2D operations. These steps remove noise and regularize region shapes, providing cleaner masks for the subsequent 3D consistency filtering.

Connected component filtering Raw prediction masks $\widehat{M}(I)$ from anomaly detection models frequently contain small, isolated false positive regions caused by reconstruction noise or model uncertainty. These typically manifest as scattered one- to few-pixel detections distributed across the background, in addition to the true metal artifact region. If left unfiltered, such spurious detections would be artificially enlarged by subsequent morphological operations. Early removal of such noise prevents the amplification of false positives.

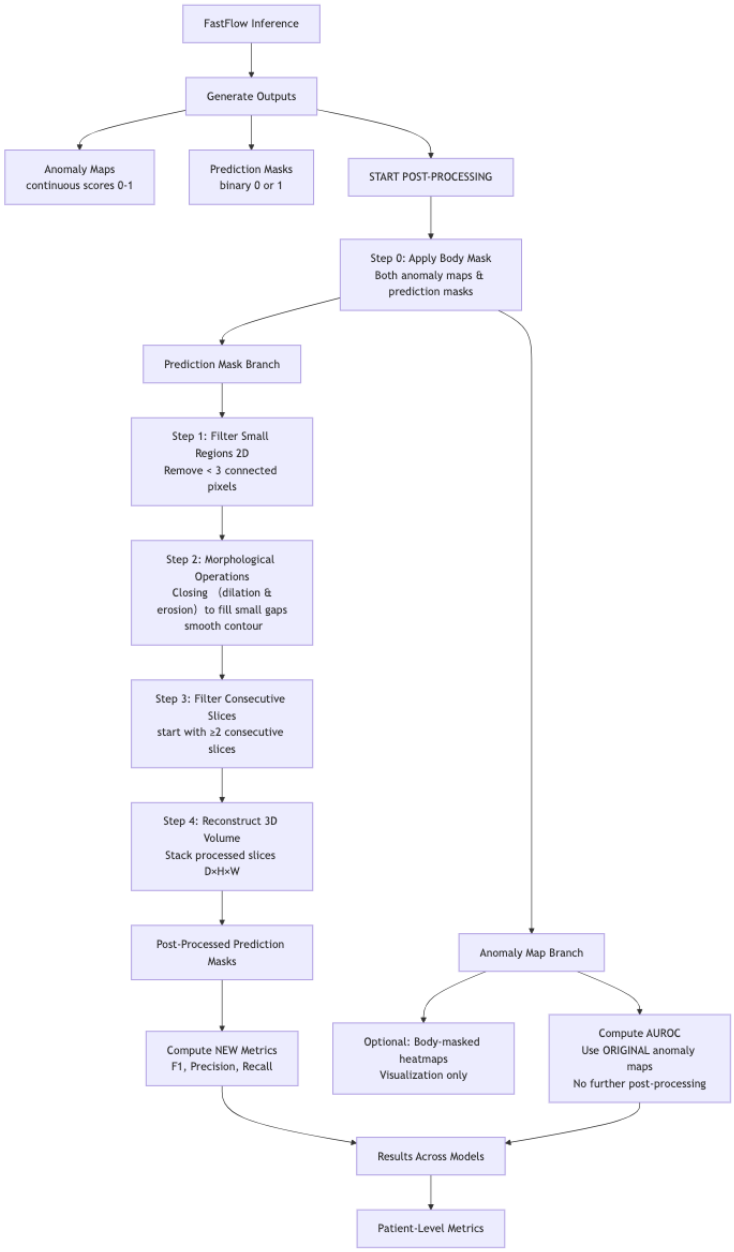


Figure 5.11: POST-PROCESSING PIPELINE OVERVIEW. *Overview of the four-stage pipeline applied to binary prediction masks prior to evaluation. Illustrated using FastFlow inference as an example, the diagram highlights the two output branches: continuous anomaly maps evaluated with threshold-independent metrics and binary prediction masks.*

Table 5.3: PIXEL CONNECTIVITY DEFINITIONS. *Comparison of 4-connectivity and 8-connectivity used for connected-component analysis in 2D images.*

Connectivity	Neighboring pixels considered
4-connectivity	North (N), South (S), East (E), West (W)
8-connectivity	North (N), South (S), East (E), West (W), North-East (NE), North-West (NW), South-East (SE), South-West (SW)

To explicitly separate coherent detections from noise, connected component analysis is performed on the raw prediction mask $\widehat{M}(I)$. This operation groups adjacent foreground pixels into spatially contiguous components, enabling filtering decisions to be made at the component level rather than at the individual pixel-level.

Connectivity between pixels is defined using 8-connectivity, in which two foreground pixels are considered connected if they share either an edge or a corner. Compared to 4-connectivity, which only considers edge-adjacent pixels, 8-connectivity better preserves irregular and diagonally extended metal artifact shapes by preventing a single coherent structure from being artificially split into multiple smaller components, which could otherwise fall below the area threshold. A schematic comparison of 4- and 8-connectivity is shown in Table 5.3.

Let Ω denote the pixel domain of image I, and let $\{C_1, \ldots, C_n\}$ be the set of connected components obtained from $\widehat{M}(I)$ under 8-connectivity. Each component $C_i \subset \Omega$ is a set of foreground pixels, with area $|C_i|$ equal to its pixel count. Components whose area falls below a minimum size threshold are discarded. The filtered mask $M_{\text{filtered}}(I) : \Omega \to \{0, 1\}$ is defined pixel-wise as

$$M_{\text{filtered}}(I, u) = \begin{cases} 1, & \text{if } \widehat{M}(I, u) = 1 \text{ and } u \in C_i \text{ with } |C_i| \geq \tau_{\text{area}}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.16}$$

where $u \in \Omega$ denotes a pixel location and $\tau_{\text{area}} = 3$ pixels is the minimum component size threshold. This choice removes 1-2 pixel noise speckles while preserving genuine small artifact fragments.

Component filtering is performed prior to morphological refinement. This design choice prevents subsequent dilation from amplifying isolated noise regions, which would otherwise necessitate more aggressive erosion for removal. The initial filtering serves as a coarse noise suppression stage, whereas subsequent morphological closing focuses on shape regularization and gap filling. Separating noise removal from shape refinement ensures that the post-processing pipeline remains interpretable and avoids redundant or conflicting operations.

Morphological Operations Following connected component filtering, the remaining prediction masks may still exhibit fragmented regions and irregular boundaries. While these artifacts are easily negligible at the pixel-level, they can lead to unstable slice-wise and patient-level decisions. For example, small gaps within an otherwise coherent artifact region may cause adjacent slices to be classified inconsistently. To improve spatial coherence and decision stability, we therefore apply morphological refinement to the binary prediction masks.

This refinement builds on the morphological framework introduced in Section 5.2.3. In contrast to the ground-truth mask generation stage, where conditional opening was used to smooth large regions while preserving small artifacts, the objective here is to connect fragmented detections and regularize boundaries. We therefore apply morphological closing, defined as dilation followed by erosion. Dilation bridges small gaps between nearby detections and reconnects fragmented parts of the same artifact, whereas erosion counteracts excessive expansion, approx-

Table 5.4: MORPHOLOGICAL PARAMETER CONFIGURATIONS. *Parameter settings for morphological closing evaluated on the validation set.*

Configuration	n_d	n_e	k	τ_{area}
Baseline	1	1	5	3
Aggressive	2	1	5	3
Conservative	1	2	5	3

imately restoring the original extent of the regions while preserving newly established connections.

A single closing operation with structuring element B of size $k \times k$ is defined as:

$$\mathcal{C}_{n_d,n_e}(M_{\text{filtered}}; B) = \left(M_{\text{filtered}} \oplus^{n_d} B\right) \ominus^{n_e} B, \tag{5.17}$$

where n_d and n_e denote the number of successive dilation and erosion iterations, respectively. The relative balance between these parameters controls the strength of refinement, trading off the ability to reconnect fragmented detections against the risk of artificial region growth.

We evaluate three simple configurations: baseline, aggressive, and conservative, each corresponding to a single closing operation with different dilation–erosion balances. All configurations use an structuring element with a kernel size k and identical connected-component filtering parameters. The evaluated parameter settings are summarized in Table 5.4.

Based on validation experiments, we adopt the aggressive configuration in all reported results. This choice favors the reconnection of fragmented artifact regions and reduces the risk of missing true artifacts.

5.5.2 3D post-processing

The slice-wise operations above treat each slice independently, but each MR scan is a 3D volume. We therefore add a volumetric post-processing stage that enforces cross-slice consistency and supports patient-wise 3D reconstruction for review.

Volumetric consistency filtering Although inference is performed on a slice-level, each MR image constitutes a 3D volume in which clinically meaningful anomalies typically persist across multiple adjacent slices. Isolated anomaly detections on a single slice are commonly caused by reconstruction noise or overly permissive thresholds. In order to reduce false positives without discarding true anomalies we enforce volumetric persistence as a final step of the post processing.

To avoid introducing additional thresholds or distance hyperparameters for 3D consistency filtering we used a simple pixelwise overlap-based persistence rule. This only keeps detections on slice z if it overlaps with either neighbor ($z-1$ or $z+1$) as detailed in the following section.

Pixelwise Overlap-based Persistence Let $M_z \in \{0,1\}^{H\times W}$ denote the refined post processed prediction mask of slice z obtained by thresholding the nomalized anomaly map at the learned operating threshold $\tau^\star$ and applying the subsequent post processing steps in 2D as discussed in Section 5.5. Since M_{z-1} and M_{z+1} are binary masks, we interpret them as foreground pixel sets where $M_z \subseteq \Omega$ with $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ contains exactly those pixels with value 1. Therefore, $|C_z^{(k)} \cap M_{z\pm1}|$ counts the number of overlapping foreground pixels. With the obvious boundary handling for $z = 1$ and $z = Z$ (use the single available neighbor). The retained mask

on slice z is the union of all components that remain:

$$M'_z = \bigcup_k C_z^{(k)}. \tag{5.18}$$

3D Volume Reconstruction After slice-wise post-processing and cross-slice consistency enforcement, we reconstruct patient-specific 3D volumes to facilitate clinical review and enable volume-based evaluation metrics. This reconstruction step serves visualization and aggregation purposes rather than influencing the detection pipeline itself.

Patient-wise stacking Slices are grouped by patient identifier (extracted from filename pattern `PA{id}_{slice}.png`) and stacked along the z-axis in anatomical order. The final 3D volume $V \in \mathbb{R}^{H\times W\times D}$ for patient p is:

$$V_p(x, y, z) = M'_{p,z}(x, y), \tag{5.19}$$

where $M'_{p,z}$ is the persistence-filtered mask for slice z of patient p, and D is the number of slices.

5.6 Evaluation

5.6.1 Slice-level Metrics

For slice-level metrics each axial slice is treated as an independent image for detection and reporting. At the same time, our post-processing enforces 3D consistency by removing anomalies that do not persist across multiple adjacent slices. This is a reasonable choice as in clinical settings radiologists review volumes slice-by-slice. Our anomaly detection pipeline produces a per-slice heatmap that is thresholded and post-processed before classification. After connected-component filtering and morphologcy to filter out any spurious regions, a slice is declared as anomalous if any post-processed anomalous region remains.

Reported metrics. At slice-level, we treat each slice as a binary classification problem: a slice after postprocessing slice is classified as positive (anomalous) if the final prediction mask contains at least one foreground pixel and negative otherwise:

We classify a slice as anomalous if any foreground pixel remains:

$$\widehat{y}^{\text{slice}}(\mathbf{I}_{i,z}) = \begin{cases} 1, & \text{if } \sum_{p\in\Omega} \widehat{M}_{\text{pp}}(\mathbf{I}_{i,z}, p) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.20}$$

The ground-truth slice label is defined analogously from the annotation mask $y(\mathbf{I}_{i,z}, p) \in \{0, 1\}$:

$$y^{\text{slice}}(\mathbf{I}_{i,z}) = \begin{cases} 1, & \text{if } \sum_{p\in\Omega} y(\mathbf{I}_{i,z}, p) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.21}$$

Slice-level confusion matrix Let $\mathcal{T}$ be the set of test slices. For each slice $(i, z) \in \mathcal{T}$, we define a predicted slice label $\widehat{y}_{i,z}^{\text{slice}} \in \{0, 1\}$ and a ground-truth slice label $y_{i,z}^{\text{slice}} \in \{0, 1\}$. The slice-level true positives TP_{slice} are the number of slices with $\widehat{y}_{i,z}^{\text{slice}} = 1$ and $y_{i,z}^{\text{slice}} = 1$, false positives FP_{slice} are those with $\widehat{y}_{i,z}^{\text{slice}} = 1$ but $y_{i,z}^{\text{slice}} = 0$, true negatives TN_{slice} are those with $\widehat{y}_{i,z}^{\text{slice}} = 0$ and $y_{i,z}^{\text{slice}} = 0$, and false negatives FN_{slice} are those with $\widehat{y}_{i,z}^{\text{slice}} = 0$ but $y_{i,z}^{\text{slice}} = 1$. These are computed over all $(i, z) \in \mathcal{T}$ and are then used to report slice-level precision, recall, F_1, and balanced accuracy:

Precision and Recall We report precision and recall on a slice-level as defined in Section 5.4. Optimizing for high recall is a conservative choice: it limits missed anomalies (false negatives) that could propagate into faulty sCT generation. However, false positives lead to real costs as this flagged MR scan then requires a CT scan. Therefore, precision must be considered in tandem; it quantifies the proportion of flagged scans who are truly anomalous. Since every false positive typically triggers additional review we want to avoid flagging many in-distribution slices while still anomalies.

F1 score In our setting we do not treat precision and recall as equally important. We therefore report precision and recall separately and include the F_1 as a compact summary of the overall trade-off. In safety-critical environments, the F_1 score may not be the optimal choice, as false negative costs may be too high. It is therefore reported next to Precision and Recall.

Balanced Accuracy The balanced accuracy (BA) is the arithmetic mean of the sensitivity (recall) and specificity:

$$S = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{5.22}$$

In comparison to plain accuracy, BA is robust under class imbalance and is therefore informative when anomalies are rare.

5.6.2 Pixel-level metrics

In anomaly detection, evaluating a model's ability to precisely localize anomalous regions is as important as measuring slice-level detection accuracy. For this purpose, pixel-level evaluation metrics are employed to assess how well the predicted prediction mask, threshold-swept metrics are not evaluated here due to complexities it would introduce and .

Since both predictions and ground truth are binary masks, evaluation is performed at a fixed threshold. Threshold-swept metrics such as AUROC can still be computed by sweeping over a set of thresholds and rerunning the post-processing. This is substantially more computationally expensive and is therefore not reported here.

Reported metrics Similar to the slice-level setting, we report **pixel-wise precision and recall**. Here, the confusion counts are computed over pixels: across all test slices $\mathbf{I} \in \mathcal{T}$ and pixels $p \in \Omega$, we count TP as the number of pixels with $\widehat{M}_{\mathrm{pp}}(\mathbf{I}, p) = 1$ and $y(\mathbf{I}, p) = 1$, FP as those with $\widehat{M}_{\mathrm{pp}}(\mathbf{I}, p) = 1$ but $y(\mathbf{I}, p) = 0$, and FN as those with $\widehat{M}_{\mathrm{pp}}(\mathbf{I}, p) = 0$ but $y(\mathbf{I}, p) = 1$. Pixel-wise precision and recall are then computed from these counts using formulas in Section 5.4. Precision measures how many predicted-positive pixels are truly positive; recall on the other hand quantifies measures how many true-positive pixels are recovered. High recall reduces missed anomalous pixels; precision limits false positive detections. (Taha and Hanbury, 2015)

Dice (pixel-level equivalent of F_1). At pixel-level, the Dice coefficient is exactly the F_1 score computed from pixel-wise precision and recall:

$$\mathrm{Dice} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = F_{1,\mathrm{pix}}. \tag{5.23}$$

Practically, "Dice" is used when interpreting the set overlap between the predicted and ground-truth masks, and "F_1" when emphasizing the precision–recall view; numerically they are identical under the same pixel-wise counts (Taha and Hanbury, 2015).

False Negative Rate (FNR). The pixel-wise miss rate (complement of recall) is computed using the confusion matrix defined above 5.6.2 and the formula defined in Section 5.4. It directly quantifies missed anomalous pixels. The pixel-level metrics above—precision/recall, Dice FNR, specificity/balanced accuracy quantify *localization*: how well predicted regions align with annotated anomalies in space.

5.6.3 Patient-level Metrics

Clinical decisions are ultimately made per scan rather than per pixel or slice. We therefore aggregate slice-level outputs into a patient-level classification.

From slices to patients Having defined image- and pixel-level metrics, the remaining question is how to aggregate slice-wise outputs into a patient-level decision that is consistent with clinical quality assurance.

There are many possible ways to aggregate slice-wise predictions into a patient-level decision, and we discuss potential improvements in the Discussion chapter Section 8.2. In this work, however, we focus on a simple and transparent aggregation rule that can be applied directly to the post-processed slices.

Let p index a patient and $z \in \{1, \dots, D_p\}$ an axial slice. After the post-processing steps described in Section Section 5.5, let $\widehat{M}_{p,z}(u) \in \{0, 1\}$ denote the final prediction mask on slice z, with pixel index $u \in \Omega$ and $\Omega = \{1, \dots, H\} \times \{1, \dots, W\}$. The number of predicted-positive pixels on slice z is

$$N^+_{p,z} = \sum_{u \in \Omega} \widehat{M}_{p,z}(u), \tag{5.24}$$

and the corresponding *positive fraction* is

$$f_{p,z} = \frac{N^+_{p,z}}{|\Omega|} \in [0, 1]. \tag{5.25}$$

Mean positive fraction. We average the slice-wise positive fractions and compare them to a threshold:

$$\bar{f}_p \;=\; \frac{1}{D_p} \sum_{z=1}^{D_p} f_{p,z}, \qquad \text{classify if } \; \bar{f}_p \geq \alpha_{\text{mean}}. \tag{5.26}$$

The statistic $\bar{f}_p$ summarizes the overall fraction of predicted-positive pixels across the scan. It reduces sensitivity to isolated false detections, but it can dilute small anomalous regions.

We use the mean positive fraction as our primary patient-level rule as it is stable, interpretable and easy to apply. The operating point is controlled by α_{mean}, and we report results for multiple values of α_{mean}. As $\alpha_{\text{mean}} \to 0^+$, the rule approaches the conservative any-positive decision.

Chapter 6

Experiments

6.1 Experiment 1: Replicated Channel Baseline

The primary goal of Experiment 1 is to establish a baseline for evaluating the anomaly detection pipeline and post-processing framework under standardized conditions on the single-center MR pelvis dataset. To ensure compatibility with anomaly detection models originally developed for RGB images, each single-channel MR slice is replicated across three input channels as described in Section 5.3. This allows the models to operate without architectural modification, enabling direct benchmarking of training stability and post-processing performance. Furthermore, the baseline provides a consistent reference point for subsequent experiments.

Two representative models are selected from each anomaly detection family introduced in Chapter 4: DRAEM and Dinomaly (reconstruction-based), STFPM and RD4AD (knowledge distillation), CFA and PatchCore (memory-bank), FastFlow and CFlow (flow-based), and DeepSVDD and CutPaste (one-class classification). Performance is assessed using the Anomalib evaluation framework[1][2]. Metrics are computed at the slice- and pixel-levels according to the definitions in Section 5.6, using Dice scores at the pixel-, slice-, and patient-levels as the primary outcome measures.

6.2 Experiment 2: Impact of Input Data Representation

Experiment 2 investigates how different input data representations affect anomaly detection performance compared to the single-slice 3-channel baseline from Experiment 1. To this end, the same single-center pelvis dataset is reformatted into two additional variants as described in Section 5.3: a *bone colormap* representation, where each MR slice is mapped to a 3-channel pseudo-RGB image with synthetic color, and a *2.5D consecutive-slices* representation, where three adjacent axial slices ($z-1$, z, $z+1$) are stacked along the channel dimension. Together with the replicated-channel baseline, these three formats allow a controlled comparison between purely replicated input, synthetic color encoding, and local volumetric context.

From the anomaly detection families introduced in Chapter 4, one high-performing configuration is selected per family based on the patient-level Dice in Experiment 1, covering reconstruction-

[1] https://anomalib.readthedocs.io/en/v2.0.0/markdown/guides/how_to/evaluation/evaluator.html

[2] https://anomalib.readthedocs.io/en/latest/markdown/guides/reference/metrics/index.html

based, knowledge-distillation, memory-bank, and flow-based methods. For each selected model, training is repeated separately on the colormap and 2.5D datasets using the same training framework as in Experiment 1. Evaluation follows the identical pipeline described in Section 5.6: raw anomaly maps are post-processed in the same way, and Dice scores at pixel-, slice-, and patient-level are reported to quantify the impact of the alternative input representations relative to the replicated-channel baseline.

6.3 Experiment 3: Backbone Replacement

Experiment 3 investigates whether replacing an ImageNet-pretrained backbone with a RadImageNet-pretrained backbone improves anomaly detection performance on pelvic MR images. To this end, we focus on FastFlow as a representative flow-based method with a clearly separated, frozen feature extractor and trainable flow head, which makes backbone replacement straightforward and computationally efficient.

In the configuration of Experiment 2, FastFlow uses a Wide ResNet-50 encoder pretrained on ImageNet. For Experiment 3, this encoder is replaced by a ResNet-50 pretrained on RadImageNet, while the network architecture and input formatting remain unchanged. As in standard FastFlow setups, the backbone remains frozen and only the normalizing-flow head is optimized during training. The dataset, input variants (replicated channels, bone colormap, and 2.5D consecutive slices), and patient-level train/validation/test split are identical to Experiments 1 and 2. Evaluation follows the same post-processing pipeline and metrics described in Section 5.6, reporting Dice scores at pixel-, slice-, and patient-level. This controlled design isolates the effect of medical-image pretraining in the backbone while keeping all other components of the anomaly detection pipeline fixed.

6.4 Experiment 4: Multi-Center Generalizability

Experiment 4 evaluates whether the most promising anomaly detection configurations from Experiment 2 maintain their performance in a multi-center setting. Specifically, for each anomaly detection family (reconstruction-based, knowledge distillation, flow-based, and memory-bank), we select the model–input-format combination that achieved the highest patient Dice scores on the single-center dataset and apply these configurations to the multi-center pelvis dataset comprising scans from centers A and C as described in Section 5.2. The training setup and evaluation pipeline are kept identical to the previous experiments, and Dice scores at pixel-, slice-, and patient-level are again reported as primary metrics. This setup allows a direct comparison of single-center versus multi-center performance for each selected configuration and isolates the impact of scanner and protocol variability on the best-performing approaches.

Chapter 7

Results

7.1 Experiment 1: Replicated Channel Baseline

Table 7.1 summarizes the postprocessed performance of all evaluated models on the single-center replicated-channel slices. Pixel-level metrics are computed from the segmentation masks, while slice-level metrics are derived from slice-wise classification. PatchCore achieves the highest Dice scores at both levels with a pixel-level Dice of 0.41 and slice-level Dice of 0.78, indicating the strongest overall segmentation and classification performance in this setting. In contrast, CFlow yields the lowest Dice scores and fails to produce meaningful pixel-level anomaly maps. The remaining reconstruction-, distillation-, and memory-bank methods fall between these extremes, with moderate localization quality and similar slice-level Dice values.

Table 7.1: EXPERIMENT 1 IMAGE- AND PIXEL-LEVEL METRICS ON REPLICATED CHANNELS. *Post-processed pixel-level and slice-level precision, recall, and Dice for all models on the single-center replicated-channel slices. Pixel-level metrics are computed from the segmentation masks, slice-level metrics from slice-wise classification. The best Dice values per level are highlighted.*

	Pixel-level			Slice-level		
Model	Prec.	Rec.	Dice	Prec.	Rec.	Dice
DRAEM	0.3522	0.4395	0.3911	0.2846	0.9855	0.4417
Dinomaly	0.0231	0.7963	0.0450	0.2392	1	0.3860
RD4AD	0.0876	0.1354	0.1063	0.3629	0.9277	0.5217
STFPM	0.0597	0.6902	0.1098	0.2393	1	0.3862
CFlow	0	0	0	0.1769	0.1253	0.1467
FastFlow	0.0404	0.0243	0.0303	0.3312	0.7422	0.4580
CFA	0.1145	0.7625	0.199	0.2402	1	0.3873
Patchcore	0.3667	0.4544	**0.4058**	0.8769	0.7036	**0.7807**

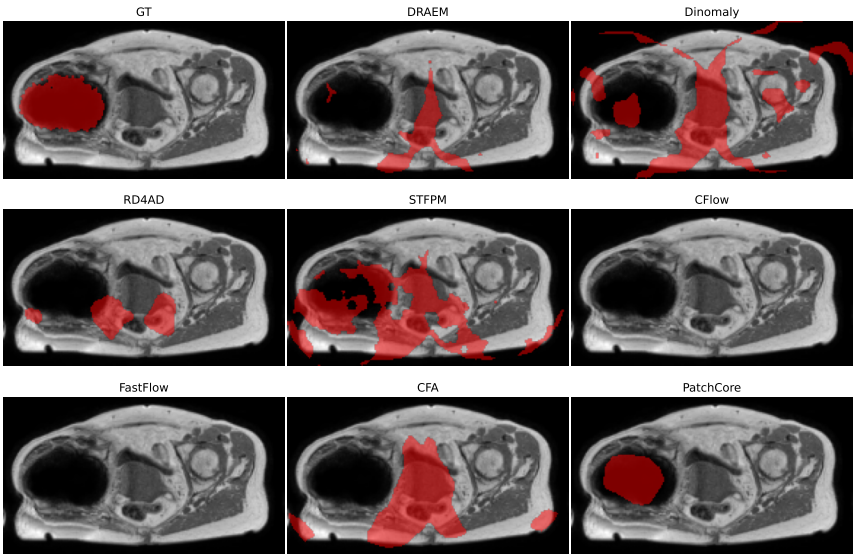
Table 7.2 reports the patient-level performance for selected decision thresholds on the replicated-channel slices. For each model, precision, recall, and Dice are listed across a small set of thresholds to characterize operating-point behavior at the patient-level without fixing a single threshold beforehand. In subsequent analyses, the best operating point is chosen as the threshold that maximizes patient-level Dice on the test set, providing an optimistic estimate of performance. Among the reconstruction-based and memory-bank methods, PatchCore reaches the highest observed patient-level Dice (0.90), followed by DRAEM and CFA, while RD4AD and STFPM show slightly

Table 7.2: EXPERIMENT 1 PATIENT-LEVEL METRICS ON REPLICATED CHANNELS. *Patient-level precision, recall, and Dice on the single-center replicated-channel slices for different decision thresholds after postprocessing. For each model, several operating points are shown without fixing a single threshold a priori. The best Dice score for each model is in bold and the best overall Dice is highlighted.*

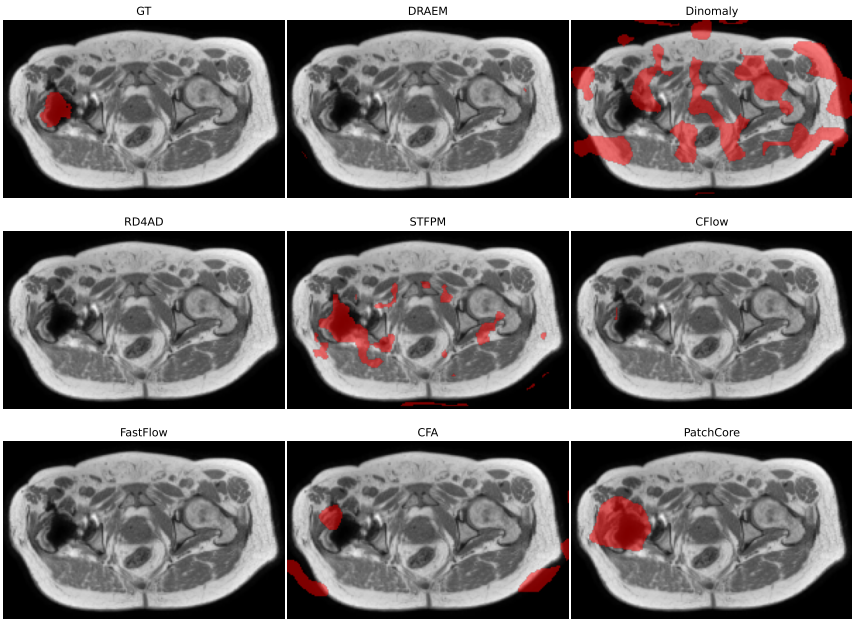
Model	Threshold	Precision	Recall	Dice
DRAEM	0	0.4667	1	0.6364
	0.002	0.8000	0.8571	**0.8276**
	0.0025	0.7692	0.7143	0.7407
Dinomaly	0	0.4667	1	0.6364
	0.025	0.5217	0.8571	**0.6486**
	0.030	0.5000	0.7857	0.6111
RD4AD	0	0.4667	1	0.6364
	0.003	0.7333	0.7857	**0.7586**
	0.0035	0.7143	0.7143	0.7143
STFPM	0	0.4667	1	0.6364
	0.030	0.9000	0.6429	**0.7500**
	0.035	0.8571	0.4286	0.5714
CFlow	0	0.5429	1	**0.7037**
	0.001	0.5000	0.0526	0.0953
	0.002	1.0000	0.0526	0.1000
FastFlow	0	0.5428	1	**0.7037**
	0.001	0.5000	0.3158	0.3871
	0.002	0.5000	0.2105	0.2963
CFA	0	0.4667	1	0.6364
	0.015	0.8	0.8571	**0.8276**
	0.02	0.8571	0.4286	0.5714
PatchCore	0	0.4667	1	0.6364
	0.0001	0.8667	0.9286	**0.8966**
	0.0005	1	0.7857	0.88

lower but still competitive best Dice values. In contrast, the flow-based models CFlow and FastFlow deteriorate rapidly once the threshold is moved away from the trivial operating point at zero.

The quantitative differences in localization quality are reflected in the qualitative examples in Figure 7.1. The OOD slice in 7.1(a) contains a large artifact area where PatchCore produces a compact anomaly region that closely overlaps the ground-truth artifact mask. Several other models either only partially cover the artifact or generate diffuse activations, with CFlow and FastFlow failing to highlight a meaningful region. The OOD slice in 7.1(b) shows a smaller artifact area, and all of the models seem to struggle more with this. While PatchCore produces an anomaly mask covering the artifact area, many models do not generate an anomaly mask at all or a false anomaly mask. For the in-distribution case in Figure 7.2, PatchCore, RD4AD, and FastFlow correctly do not produce an anomaly mask, whereas some reconstruction- and flow-based models still predict spurious anomalous areas despite the absence of metal artifacts.



(a) OOD volume 1PA118 slice 57.



(b) OOD volume 1PA147 slice 51.

Figure 7.1: EXPERIMENT 1 PREDICTED ANOMALY MASKS OF THE EVALUATED MODELS FOR OOD CASES. *Visual comparison of anomaly masks (red overlay) generated by the evaluated models against the ground-truth (GT) artifact mask for two different OOD volumes.*

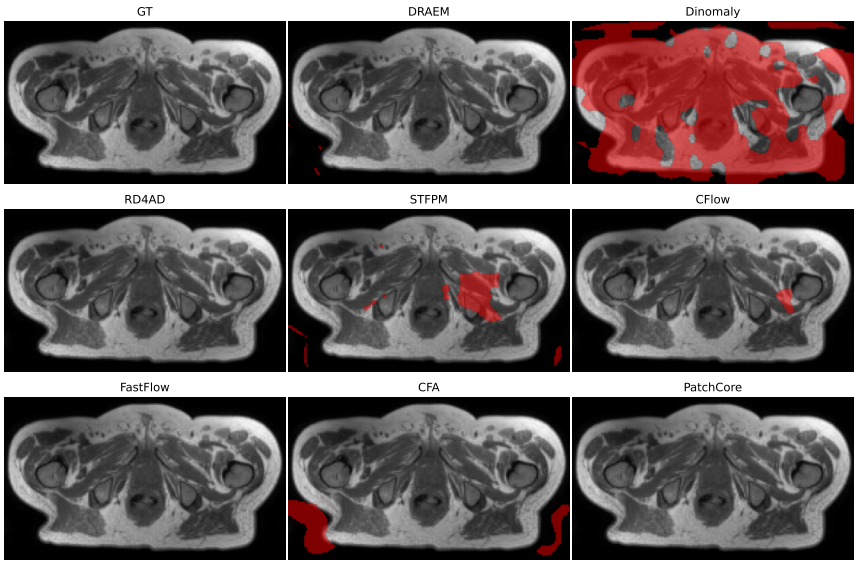
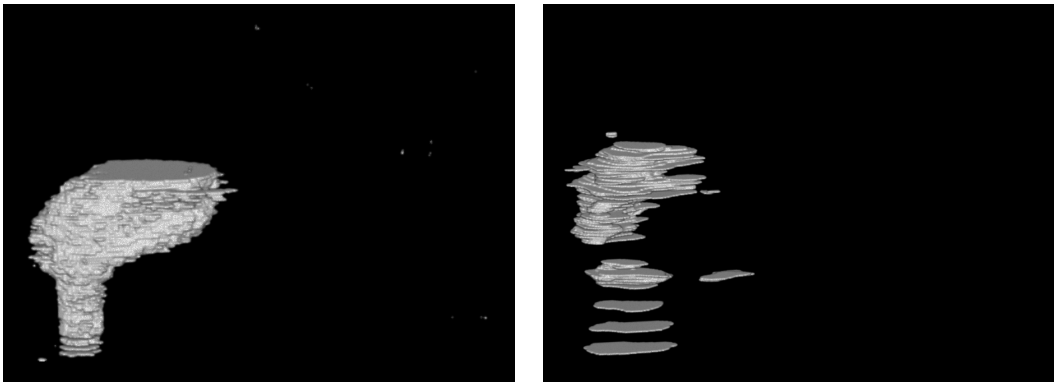


Figure 7.2: EXPERIMENT 1 PREDICTED ANOMALY MASKS OF THE EVALUATED MODELS FOR IN-DISTRIBUTION CASE.. *Anomaly masks (red overlay) generated by the evaluated models against the ground-truth (GT) artifact mask for the OOD volume 1PA058 (slice 45).*

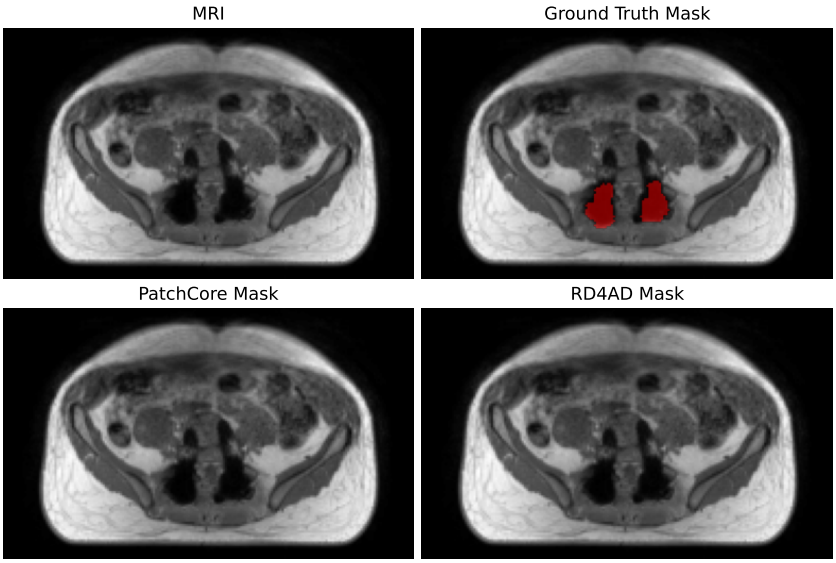


(a) Ground-truth artifact volume. (b) PatchCore anomaly volume.

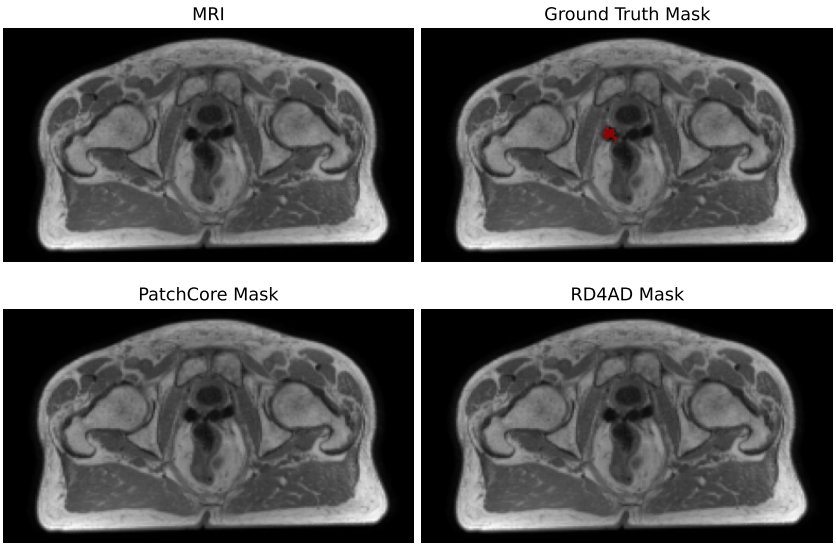
Figure 7.3: EXPERIMENT 1 VOLUMETRIC BEHAVIOR OF PATCHCORE. *Three-dimensional visualization of the ground-truth artifact volume and the corresponding PatchCore anomaly volume for the OOD case 1PA118.*

The volumetric comparison in Figure 7.3 shows that the ground truth artifact forms a coherent three-dimensional region around the implant and that the PatchCore anomaly volume follows this structure. This further supports PatchCore as the strongest contender in this baseline setting.

In pelvis MR, very small metal objects often create only tiny signal voids or subtle intensity changes that resemble other low-signal or high-contrast structures, such as air pockets or bowel gas. The examples in Figure 7.4 illustrate that even the strong models fail to detect such subtle artifacts: none of the evaluated methods produces an anomaly mask on the relevant slices, despite



(a) Visualization of the anomaly mask prediction for volume 1PA136 slice 110.



(b) Visualization of the anomaly mask prediction for volume 1PA152 slice 63.

Figure 7.4: EXPERIMENT 1 CHALLENGING SMALL-IMPLANT CASES. *Examples of small implants with subtle artifacts. The ground truth and anomaly masks are marked red on top of the MR slice. Both models in (a) and (b) fail to provide a anomaly mask for this slice.*

Table 7.3: EXPERIMENT 1 PERFORMANCE BEFORE AND AFTER POSTPROCESSING. *Comparison of model performance on the single-center replicated-channel slices before and after the postprocessing pipeline. AUROC is reported only for the raw anomaly scores (pixel- and slice-level) before postprocessing, while Dice scores are shown at pixel-, slice-, and patient-level for both configurations. The best patient-level Dice corresponds to the threshold that maximizes patient Dice on the test set.*

Model	AUROC Pixel / Image	Pixel Dice Before	Pixel Dice After	Slice Dice Before	Slice Dice After	Best Patient Dice Before	Best Patient Dice After
DRAEM	0.9618 / 0.8655	0.3504	↑ 0.3911	0.3860	↑ 0.4417	0.8387	↓ 0.8276
Dinomaly	0.6269 / 0.3160	0.0071	↑ 0.0450	0.3860	= 0.3860	0.6364	↑ 0.6486
RD4AD	0.9539 / 0.8002	0.0857	↑ 0.1063	0.5226	↓ 0.5217	0.7586	= 0.7586
STFPM	**0.9632** / 0.7380	0.0907	↑ 0.1098	0.3861	↑ 0.3862	0.7143	↑ 0.7500
CFlow	0.5703 / 0.5279	0.0000	= 0.0000	0.1657	↓ 0.1467	0.7037	= 0.7037
FastFlow	0.9154 / 0.6591	0.0271	↑ 0.0303	0.4723	↓ 0.4580	0.7037	= 0.7037
CFA	0.9692/ 0.7498	0.0799	↑ 0.199	0.4947	↓ 0.3873	0.7059	↑ 0.8276
PatchCore	0.9298 / **0.9821**	0.4052	↑ 0.4058	0.7807	= 0.7807	0.8966	= 0.8966
DeepSVDD	– / 0.8075	–	–	0.5982	–	–	–
CutPaste	– / 0.7498	–	–	0.4947	–	–	–

the ground-truth indicating small artifact regions. This highlights that small-implant cases remain outside the effective detection range of the current models and post-processing pipeline.

Table 7.3 compares model performance on the replicated-channel slices before and after the postprocessing pipeline. AUROC is reported only for the raw anomaly scores before postprocessing, while Dice is provided at pixel-, slice-, and patient-level for both configurations. STFPM attains the highest pixel-level AUROC (0.96), whereas PatchCore achieves the highest slice-level AUROC (0.98). CFlow and Dinomaly form the lower end of the spectrum with markedly reduced AUROC at both pixel- and slice-level. Across models, postprocessing typically increases pixel-level Dice by roughly 10–150% relative to the preprocessed values, while slice- and patient-level Dice remain unchanged or change only modestly. CFA exhibits one of the strongest relative gains in pixel-level Dice after postprocessing, whereas PatchCore and the flow-based models are largely unaffected in terms of their best patient-level Dice.

DeepSVDD and CutPaste are only reported at slice-level in this comparison and do not provide pixel-wise localization maps in the current setup, which prevents the use of the segmentation-oriented postprocessing and the computation of pixel- or patient-level Dice for these models.

7.2 Experiment 2: Impact of Input Data Representation

Table 7.4 summarizes the postprocessed performance of the best-performing models from Experiment 1 when applied to the three alternative input formats in this experiment. For each combination of model and format, pixel-level Dice, slice-level Dice, and the best patient-level Dice over evaluated thresholds are reported, allowing a direct comparison between replicated channels, adjacent slices, and bone colormap inputs on the same test set.

Table 7.4: EXPERIMENT 2 METRICS FOR DIFFERENT INPUT FORMATS. *Postprocessed performance of selected models for three input formats on the same test set: replicated channels, adjacent slices, and bone colormap. For each model–format combination, pixel-level Dice, slice-level (image) Dice, and the best patient-level Dice over evaluated thresholds are reported.*

Model	Input format	Pixel Dice	Slice Dice	Best Patient Dice
DRAEM	Replicated channels	**0.3911**	**0.4417**	**0.8276**
	Adjacent slices	0.1157	0.3860	0.7097
	Bone colormap	0.0803	0.3860	0.7222
RD4AD	Replicated channels	0.1063	**0.5217**	**0.7586**
	Adjacent slices	0.0882	0.5181	**0.7586**
	Bone colormap	**0.1432**	0.4665	0.6471
FastFlow	Replicated channels	**0.0303**	**0.4580**	**0.7307**
	Adjacent slices	0.0030	0.3965	0.7317
	Bone colormap	0.0005	0.4519	0.6857
PatchCore	Replicated channels	0.4058	**0.7807**	0.8966
	Adjacent slices	0.4050	0.7750	0.8800
	Bone colormap	**0.5097**	0.7682	**0.9630**

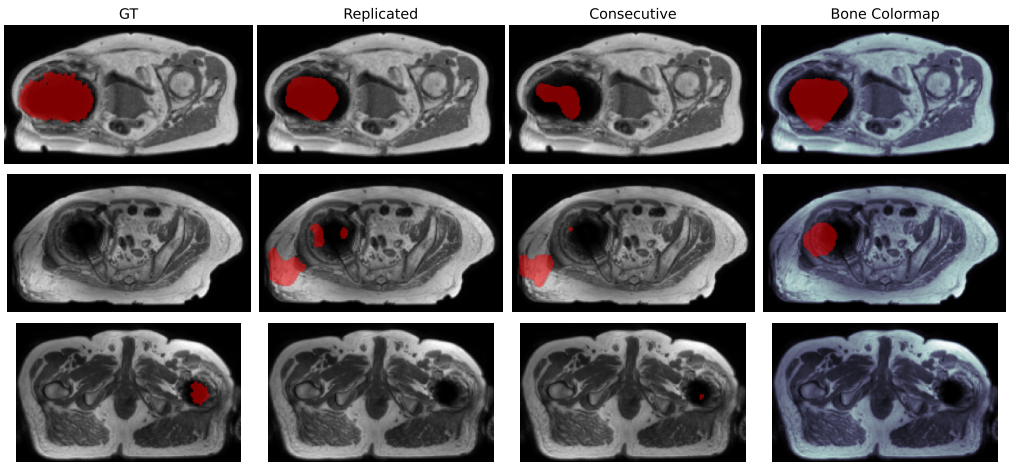


Figure 7.5: EXPERIMENT 2 PATCHCORE ANOMALY MASKS FOR DIFFERENT INPUT FORMATS. *Predicted anomaly masks (red overlay) from the PatchCore model for three OOD volumes 1PA118 (slice 57), 1PA155 (slice 112), and 1PA169 (slice 65) using replicated channels, adjacent slices, and bone colormap inputs.*

Across formats, PatchCore remains the strongest method. Its slice-level Dice stays consistently high around 0.78 for all three input formats, while pixel-level Dice increases from 0.41 for replicated channels to 0.51 for the bone colormap, and the best patient-level Dice rises from 0.90 to 0.96, indicating that the colormap representation improves both localization and patient-level decisions. These trends are illustrated in Figure 7.5. For representative OOD slices, the PatchCore model with replicated channels produces a compact anomaly mask around the artifact, while the colormap-based model covers the artifact region more completely. RD4AD also benefits from

Table 7.5: EXPERIMENT 3 FASTFLOW BACKBONE COMPARISON. *Postprocessed performance of FastFlow with ImageNet and RadImageNet backbones across three input formats. For each backbone–format combination, pixel-level Dice, slice-level Dice, and the best patient-level Dice over evaluated thresholds are reported.*

Backbone	Input format	Pixel Dice	Slice Dice	Best Patient Dice
ImageNet	Replicated channels	0.0303	**0.4580**	0.7037
	Adjacent slices	0.0030	0.3965	0.7317
	Bone colormap	0.0005	**0.4519**	0.6857
RadImageNet	Replicated channels	**0.0327**	0.4500	**0.7317**
	Adjacent slices	**0.0796**	**0.4319**	**0.7907**
	Bone colormap	**0.0457**	0.4404	**0.7222**

the bone colormap at the pixel-level but maintains broadly similar slice- and patient-level Dice for replicated channels and adjacent slices. In contrast, DRAEM and FastFlow achieve their best pixel-level Dice with replicated channels and show clearly weaker localization with adjacent slices and colormap inputs, while their slice- and patient-level Dice vary only modestly between formats and remain below those of RD4AD and PatchCore.

7.3 Experiment 3: Backbone Replacement

Table 7.5 compares the post-processed performance of FastFlow when using a standard ImageNet-pretrained backbone versus a RadImageNet-pretrained backbone across the three input formats from Experiment 2. For each backbone–format combination, pixel-level Dice, slice-level Dice, and the best patient-level Dice over evaluated thresholds are reported, enabling a direct assessment of whether domain-specific pretraining improves FastFlow in this setting.

Overall, replacing ImageNet with RadImageNet slightly improves FastFlow but does not resolve its core limitations. Slice-level Dice remains in a similar range for both backbones, while RadImageNet yields consistently higher pixel-level Dice, corresponding to relative gains on the order of 50–90% across the three input formats, and modest increases in best patient-level Dice of roughly 3–8% depending on the format. Nevertheless, the absolute Dice values stay clearly below those of the strongest methods from Experiments 1 and 2. This is also reflected in the example in Figure 7.6, where FastFlow's anomaly maps remain fragmented and noisy compared to the more compact and accurate masks produced by PatchCore (as seen in Figure 7.1).

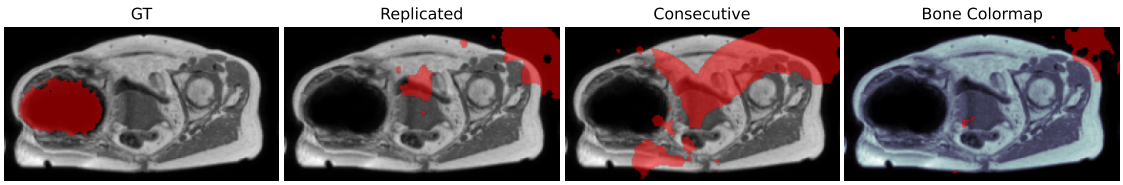


Figure 7.6: EXPERIMENT 3 FASTFLOW ANOMALY MASK PREDICTIONS WITH RADIMAGENET BACKBONE. *Predicted anomaly maps (red overlay) for the OOD volume 1PA118 (slice 57) using the FastFlow model with the RadImageNet backbone and the 3 different input formats.*

7.4 Experiment 4: Multi-Center Generalizability

Table 7.6 summarizes how the performance of the best-performing models changes when moving from the single-center to the multi-center replicated-channel datasets. For each model and dataset, pixel-level Dice, slice-level Dice, and the best patient-level Dice over evaluated thresholds are reported, allowing a direct assessment of robustness to increased acquisition variability.

Overall, all methods experience a performance drop on the multi-center data, but the magnitude differs across models and metrics. Pixel- and patient-level Dice for DRAEM, RD4AD, and FastFlow typically decrease by about 30–40%, with notable reductions in slice-level Dice as well, indicating reduced robustness of these models under protocol and scanner heterogeneity. PatchCore also degrades on the multi-center bone colormap data, but its relative drop is smaller (around 5% at pixel-level and 20–25% at slice- and patient-level), and it continues to achieve the highest Dice scores among the evaluated methods in this more challenging setting.

Table 7.6: EXPERIMENT 4 SINGLE- VS. MULTI-CENTER PERFORMANCE. *Dice performance of the best-performing model-format combinations when moving from a single-center dataset to a multi-center dataset. For each model and dataset, pixel-level Dice, slice-level Dice, and the best patient-level Dice over evaluated thresholds are reported.*

Model	Dataset	Pixel Dice	Slice Dice	Best Patient Dice
DRAEM	Single-center replicated	0.3911	0.4417	0.8276
	Multi-center replicated	0.1832	0.2925	0.5926
RD4AD	Single-center replicated	0.1063	0.5217	0.7586
	Multi-center replicated	0.0318	0.4034	0.6452
FastFlow	Single-center replicated	0.0030	0.3965	0.7317
	Multi-center replicated	0.0009	0.2868	0.5614
PatchCore	Single-center bone colormap	**0.5097**	**0.7682**	**0.9643**
	Multi-center bone colormap	0.4856	0.7099	0.7742

Chapter 8

Discussion

8.1 Discussion of Experimental Findings

The baseline **Experiment 1** on single-center replicated-channel slices suggests that methods exploiting local feature memories are better aligned with metal-implant artifacts than reconstruction- or flow-based approaches. PatchCore's patch-wise memory and nearest-neighbor scoring plausibly favor compact, localized deviations, which match the way metal artifacts appear as structured, confined regions and may explain its consistently higher Dice and more coherent masks. PatchCore also appears able to leverage a feature extractor pretrained on natural images (ImageNet) and still transfer it reasonably well to pelvic MR, despite the domain mismatch. Recent work has shown that PatchCore-style methods remain competitive when applied to medical imaging benchmarks, supporting the hypothesis that memory-based patch features can tolerate a backbone trained on non-medical data and still generalize beyond the industrial texture settings for which they were originally proposed (Barusco et al., 2025; Roth et al., 2022). Other methods, such as the flow-based models, struggle more with this domain shift and show weak pixel-level performance. This might reflect that likelihood modeling in a feature space pretrained on natural images is brittle once the underlying MR feature distribution moves away from the source domain, in line with recent reports that many OOD detectors fail to generalize reliably in 3D medical imaging settings (Vasiliuk et al., 2023). Reconstruction and teacher–student models sit between these extremes. They often detect an anomaly on a slice, but either over-smooth subtle artifacts or respond to normal high-contrast structures, leading to diffuse or fragmented masks. Pelvic MR anatomy is highly variable, and normal structures such as bowel gas or bony interfaces can resemble metal-induced voids. In such settings, teacher–student models seem to struggle, because the student is encouraged to mimic the teacher on appearances that sometimes look 'artifact-like' even when they are normal, which can blur the distinction between true metal artifacts and challenging but normal anatomy and contribute to coarser localization.

The results also underline that AUROC alone seems to be a poor proxy for clinical utility. Several models achieve high AUROC but still produce misaligned or systematically too small or too large masks, which reduces Dice and would complicate patient-level decisions in practice. This pattern is consistent with broader concerns in medical imaging that strong classification AUROC does not guarantee reliable localization or clinically meaningful performance (Kocak et al., 2025).

The modest and method-dependent impact of the postprocessing pipeline suggests that it can refine but not fundamentally repair anomaly maps. When a reasonable structure is present, simple smoothing and component filtering improve pixel-level Dice slightly, but when maps are noisy or nearly empty, the same pipeline cannot hallucinate a meaningful artifact region. Together with the qualitative failures on small-implant cases, where all models miss subtle artifacts, this

indicates that the current combination of backbone, training objective, and generic postprocessing is biased toward larger, more conspicuous artifacts.

Experiment 2 investigated how changes in input representation affected the performance of the models. The results show that the input format interacts with the model families in subtle ways, and the small numerical differences observed should be interpreted cautiously in light of the general limitations. For PatchCore, the slight advantage of the bone colormap over replicated channels is at least coherent with its design: a memory over precomputed feature patches should, in principle, benefit from a richer channel structure that makes local patterns more separable in feature space, without altering the underlying anatomy. For RD4AD, the mixed picture fits the idea that a teacher trained on natural images may not fully exploit synthetic color in MR. For reconstruction- and likelihood-based methods such as DRAEM and FastFlow, the absence of clear benefits from 2.5D or colormap inputs may indicate that these architectures are more tightly coupled to specific assumptions about intensity distributions and channel semantics, making them less flexible to changes in representation. At the same time, many of the observed differences between formats are small and within the range that could plausibly arise from stochastic effects in training. Overall, these results are consistent with prior work highlighting that the benefits of ImageNet pretraining and the robustness of OOD methods in medical imaging are highly sensitive to domain and representation mismatch (Raghu et al., 2019; Vasiliuk et al., 2023).

Experiment 3 probed whether a medical-image–pretrained backbone can compensate for the domain gap between natural images and pelvic MR in FastFlow. Replacing ImageNet with RadImageNet consistently increases pixel-level Dice and improves patient-level Dice across all three input formats, with the largest gain for the adjacent-slice setting, which is plausibly explained by the encoder having been pretrained on radiological images instead of natural photographs and therefore providing features that better match MR appearance. This pattern is coherent with broader evidence that feature-space anomaly detectors are sensitive to the choice of pretrained backbone and tend to benefit when the pretraining task is better aligned with the target domain (Heckler et al., 2023). Slice-level Dice, however, remains largely unchanged, suggesting that the per-slice decision is still predominantly governed by the normalizing-flow component and is less sensitive to refinements in the encoder representation. Overall, the impact of RadImageNet pretraining on FastFlow is therefore mixed. The backbone provides more informative features for voxelwise localization, but these gains do not consistently carry over to slice- or patient-level metrics. This could be because FastFlow operates on already downsampled backbone features, so some of the added detail from a better encoder is discarded before reaching the flow head. Limitations for this experiment are that RadImageNet is trained jointly on heterogeneous, multi-center CT, MR, and ultrasound data from multiple body regions, which narrows but does not remove the domain gap to the much narrower setting of single-center pelvic MR images in this experiment (Mei et al., 2022). Additionally, only the backbone in one flow-based architecture is modified, so the findings should be viewed as model-specific for FastFlow rather than a general verdict on RadImageNet, as other families, such as distillation or memory-bank methods, may interact differently with medical pretraining and potentially benefit more or less.

Our multi-center analysis in **Experiment 4** shows that performance generally degrades when moving from a single-center to a multi-center setting, confirming that scanner- and protocol-induced variability remains a key obstacle for robust deployment. The larger reductions at slice- and patient-level metrics suggest that it becomes harder to turn local anomaly evidence into stable per-scan decisions once acquisition heterogeneity increases. This observation is consistent with broader evidence that deep learning models for MR imaging often suffer marked performance drops under scanner domain shift and cross-site deployment, underscoring the need for explicit strategies to handle MR heterogeneity (Guo et al., 2024; Ottoni et al., 2025). Overall, the experiment suggests that with the current multi-center sample size, none of the models achieves fully robust multi-center generalization.

Several limitations qualify these findings across all experiments. The analysis is restricted to a small set of neural-network architectures from reconstruction, distillation, flow, and memory-bank families, so the observed robustness patterns may not transfer to other designs or future methods. The available scanners and imaging protocols are determined by the SynthRAD2023 challenge data rather than by a prospectively designed sampling of vendors, anatomies, and sequence types, which may underrepresent important sources of real-world variability. For Experiments 1–3, the use of a single-center pelvis cohort limits conclusions about generalization to other body regions and scanners. In addition, the OOD phenomena studied here are dominated by a single artifact type (metal implants), which constrains how far the results can be extrapolated to other sources of OOD signals. Ground-truth masks are derived from 2D CT-based delineations of metal implants rather than direct MR artifact contours, so MR artifact regions that extend beyond the implant volume may be partially unlabeled, leading to underestimation of localization performance for models that correctly highlight those areas. Training hyperparameters such as learning rate, number of epochs, and batch size were explored only in a limited fashion, so the reported models may not represent the best achievable configurations for the chosen architectures. In addition, patient-level decision thresholds were selected on the test set rather than fixed on a separate validation set, which makes the corresponding Dice scores optimistic estimates and reduces their value as unbiased performance indicators. Moreover, no explicit MR harmonization or intensity-normalization strategies were applied before training and evaluation, even though such preprocessing could plausibly improve cross-center stability.

8.2 Future Work and Improvements

Building on the limitations and failure modes discussed above, several directions for improving data curation, model design, and evaluation emerge. Closer collaboration with clinical experts to create refined MR-based artifact masks could better capture subtle but clinically relevant artifact regions, potentially focusing first on smaller regions of interest around implants where manual labeling is tractable, although the feasibility of this improvement will depend strongly on available resources. Another direction would be to explore three-dimensional pre-processing and artifact detection, for example, by leveraging full volumes rather than single slices, to obtain more consistent anomaly regions across neighboring slices and to better capture the true 3D extent of metal-induced signal loss.

Beyond improved pre-processing and annotations, future work on the model side could explore architectures that better exploit three-dimensional context and larger training sets. 3D or hybrid 2.5D anomaly-detection models that process full volumes or stacks of slices may capture the spatial extent of metal artifacts more reliably and could reduce inconsistent predictions across neighboring slices. In addition, training the most promising approaches, such as memory-bank and distillation models, on larger and more diverse multi-center cohorts would help them learn a broader in-distribution representation and improve robustness to scanner and protocol variability. Finally, combining medical-image pretraining (e.g., RadImageNet-style backbones) with explicit harmonization or domain-adaptation strategies tailored to MR could further reduce the domain gap between training and deployment settings and should be evaluated in prospective, clinically realistic MR-only radiotherapy workflows.

Beyond the choice of model and the input preprocessing, the post-processing pipeline is another component where there remains room for improvement.

In this work, we adopt a simple persistence rule to enforce cross-slice consistency in binary prediction masks. Alternative strategies could be explored in future studies. For instance, connected components across adjacent slices could be matched by comparing the Euclidean distances between their centroids. Such an approach would require defining a distance threshold to deter-

mine whether an anomaly persists across slices. While straightforward, this method presents several challenges: irregular or small regions may exhibit large centroid shifts despite corresponding to the same structure, and identifying a single threshold that generalizes across different anatomies and spatial scales is non-trivial. Another option to explore is to require sufficient volumetric overlap of matched components like a Dice threshold. This approach is more shape-aware but introduces new hyperparameters and is sensitive to boundary fluctuations after morphology. Small but genuine anomalies may fail a strict overlap threshold, while permissive thresholds may admit many false positives. These considerations motivated our choice of a different persistence criterion, though centroid-based matching and volumetric overlap remain a potential direction for future investigation.

In anomaly detection and medical imaging in particular the area under the ROC curve (AUROC) is one of the most commonly reported summary metrics: models produce continuous anomaly scores and the AUROC captures how well the true positive rate performs under a given false positive rate for all thresholds. In our anomaly detection pipeline, however, the deployed evaluation does not use raw anomaly scores: the continuous pixel scores are thresholded at a first level τ to obtain a binary prediction mask, and only then do we apply connected-component filtering and morphological opening/closing. The reason is structural: these operators are defined on sets of foreground pixels (binary images), where "objects" and "background" are not ambiguous. This leads to concrete semantics: a region is flagged iff it survives thresholding and satisfies spatial rules.

This practically yields us three benefits: **(i)** *Object definition:* connected-component labeling requires a foreground/background predicate to identify objects under 8-neighbor connectivity. **(ii)** *Spatial priors:* working in the binary domain lets us impose clear spatial rules. Morphological opening removes isolated, very small detections; morphological closing bridges gaps that are shorter than a chosen neighborhood size. A simple area threshold enforces a minimum region size. These steps reduce speckle-like false positives while preserving clinically meaningful anomalies. **(iii)** *radiology review:* reviewers operate in terms of concrete anomalous regions, not continuous score fields; binary masks map can therefore directly be using in their workflow.

In our setting, post-processing is applied at a the learned optimal threshold. While an AUROC can still be approximated by sweeping over multiple input thresholds and rerunning the full post-processing for each value. We do not report it here because computing an ROC curve would require sweeping over a set of threshold and rerunning the full post-processing pipeline for each threshold. This would make the AUROC depend on a chosen threshold grid and would substantially increase computational cost, so we focus on fixed-threshold metrics on the post-processed prediction masks.

Having defined how we evaluate post-processed slice predictions, the next question is how to aggregate slice-level outputs into a single patient-level decision. In our work, we adopted a simple aggregation method, however, there are further options that could be explored in future works. One option is to classify a 3D volume using the estimated anomalous volume V_p^+. The 3D volume is classified as anomalous if $V_p^+ \geq V_{\min}$ (mm^3). This enforces a minimum 3D volume for an anomaly and which makes patient-level aggregation robust to isolated speckles, but choosing a meaningful $V_{\min}$ is non-trivial. A related alternative is to introduce a **global positive fraction**

$$F_p = \frac{\sum_z N_{p,z}^+}{\sum_z N_{p,z}}, \qquad \text{classify if } F_p \geq \alpha_{\text{vol}}. \tag{8.1}$$

This approach aggregates all predicted anomalous pixels across the scan and normalizes them by the total number of pixels, which makes the decision less sensitive to different patient scan sizes. A drawback, however, is that small but clinically relevant anomalies can be diluted in large volumes. A conservative alternative is to classify a volume as anomalous if any of its slices

contains predicted anomalous pixels:

$$y_p = \begin{cases} 1, & \text{if } \sum_{z=1}^{Z_p} \sum_{p' \in \Omega_{p,z}} \hat{M}_{p,z}(p') > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8.2}$$

Equivalently, $y_p = 1$ if $\max_z N^+_{p,z} > 0$ (or $\max_z f_{p,z} > 0$). The advantage of the any positive voxel approach is that it reduces missed anomalies at a scan-level. However, it can also be triggered by a single false-positive slice potentially making this measure overly sensitive.

Another alternative could have been to use Multi instance learning with attention aggregation. Wang et al. (2025) present a patient-level diagnoses framework for MR imaging where each exam is treated as a bag containing many slice-level instances. Instead of evaluating slices independently each MR scan is treated as a bag of slices and attention-based multiple instance learning is used to aggregate slice information into a single patient-level prediction. This is commonly implemented by a Dual-Encoder which extracts instance features, and a Multi instance learning (MIL) classifier aggregates them at a bag level. A Siamese Teacher-Student learns attention weights that highlight potentially diagnostically informative patterns and suppress noise.

In practice, radiologists go analyze a MR scan by giving more weight to slices with the most diagnostic evidence. Attention-style MIL aggregation mirrors this workflow by learning which instances matter most across an entire MR scan. This might lead to the most clinically relevant patient-level decision.

However, Wang et al. (2025) operate in a fully supervised tumor-classification setting with sufficient labels. Their pipeline requires supervised learning of both the instance coder and bag-level aggregator and they themselves note data-scale constraints. Our current anomaly detection pipelines isquite different: In constrast to supervised frameworks, our pipleine is fully unsupervised. This is clinically relevant since high-quality radiology labels are time-consuming, costly and often inconsistent across scanners and sequences. Furthermore, given the limited labels available, methods that require supervised patient-level aggregation are impractical at present. Lastly, the concept of Cross-/dimensional transfer could have been used for patient-level aggregation. Messaoudi et al. (2023) explore how to leverage 2D networks for volumetric (3D) medical image segmentation by (i) **weight transfer**—embedding a pre-trained 2D encoder inside a higher-dimensional U-Net (ii) **dimensional transfer** expanding 2D weights to initialize a 3D encoder ("Dimensionally-eXpanded" DX-Net) They define cross-dimensional transfer as expanding 2D weights to initalize 3D models. This is followed by the validate the idea across MR/CT/US benchmarks (CAMUS, CHAOS, BraTS), reporting competitive Dice scores. In practice, 6their 3D encoders are initialized from Imagenet-pretrained 2D EfficientNet weights (noisy Student) and then supervised on segmentation labels to learn volumetric anatomy and pathology (Messaoudi et al., 2023). Cross-/dimensional transfer offers a data-efficient way to capture 3D information through slice context by expanding 2D backbones to initalize 3D models. This could improve case-level robustness (Messaoudi et al., 2023). However, this route assumes supervised training with voxel-level annotations and substantial compute resources. In contrsat our pipeline is unsupervised and operates on post-processed prediction masks. Only few labels are available. Training an end-to-end 3D segmentation is therefore not feasible at present. For now, we obtain patient-level decisions via simple auditable aggregation of slices.

Chapter 9

Conclusion

This study addresses the challenge of detecting OOD cases in MR imaging for MR-to-sCT workflows, with a focus on a single, well-defined implant-metal OOD category. A ground truth mask generation pipeline was constructed that leverages the high-density metal signal in CT through adaptive thresholding, followed by mask refinement in the MR domain using MR-guided flood fill and conditional morphological processing. Ten representative anomaly detection methods from five model families were then evaluated, following the taxonomy used in BMAD: reconstruction-based, one-class classification, knowledge distillation, memory-bank, and flow-based approaches. All experiments used the same post-processing pipeline, which improved pixel-level localization across methods. Four experiments examined baseline performance on single-center data, the impact of input representations, domain-specific pretraining with RadImageNet, and multi-center generalizability.

On single-center data, PatchCore achieved the highest overall performance, with a pixel-level Dice score of 0.41, a slice-level Dice score of 0.78, and a patient-level Dice score of 0.90 using replicated-channel input. Among input representations, bone colormap encoding improved PatchCore, yielding a pixel-level Dice score of 0.51 and a patient-level Dice score of 0.96, the best results across all experiments. Incorporating adjacent-slice context did not yield consistent benefits and, in several models, reduced localization accuracy. Flow-based methods exhibited unstable behavior. Replacing ImageNet with RadImageNet pretraining led to relative gains of approximately 50–90% in pixel-level Dice scores for FastFlow, but the corresponding improvements in patient-level Dice remained modest, and FastFlow still did not close the performance gap to the stronger methods. Multi-center evaluation indicated that all methods experienced performance declines due to increased acquisition variability. PatchCore demonstrated the greatest robustness, with only a 5% reduction at the pixel-level and 20–25% reductions at the slice- and patient-levels, compared to 30–40% declines observed in other methods. However, all evaluated methods failed to detect very small implants that produced only subtle signal voids, highlighting a systematic limitation when artifact size falls below the effective detection threshold.

In summary, this study presents a systematic evaluation of anomaly detection methods for implant-metal artifacts in pelvic MR imaging. Both pre- and post-processing pipelines and the evaluation framework were designed to enable controlled and transparent comparisons under practical constraints. The observed failure cases suggest that developing enhanced approaches to detect subtle artifacts and achieve robust cross-center generalization remains an open challenge.

Chapter 10

Contributions

This master's project represents a collaborative effort of all team members. In this section, we highlight our individual contributions to the project.

- **Ömer Yildirim**
  - **Project Coordination:** Maintained structured documentation of meeting outcomes, decisions, and deadlines.
  - **Literature Review:** Conducted the review on synthetic CT generation and PSQA for MR-only radiotherapy, forming the content of the Related Work chapter.
  - **Annotation Strategy:** Documented the annotation strategy for the SynthRAD23 dataset to standardize the labeling rules for the team.
  - **Dataset Annotation:** Performed manual annotations on the SynthRAD23 dataset.
  - **Reconstruction-Based Methods:** Investigated reconstruction-based AD families, implementing architectures such as DRAEM, DSR, and Dinomaly.
  - **Experiments with DRAEM and Dinomaly:** Executed the experiments from model training to custom post-processing for the DRAEM and Dinomaly models.
  - **Sections & Chapters:** Wrote the Introduction and Related Work chapters, as well as Magnetic Resonance (MR) Imaging, Reconstruction-Based Anomaly Detection, and Dataset Annotation sections.
- **Yeyang Liu**
  - **Dataset Annotation:** Performed manual annotations on the SynthRAD23 dataset.
  - **Dataset Preprocessing Pipeline:** Analyzed the dataset and explored techniques for preprocessing. Developed a slice-level OOD dataset construction pipeline from NIfTI files, including initial iterations of refinements, candidate mask generation, and NIfTI format support.
  - **Memory Bank Methods**: Investigated memory bank-based AD approaches, implementing architectures such as PaDiM, CFA, and PatchCore.
  - **Experiments with CFA and PatchCore**: Executed the experiments from model training to custom post-processing for the CFA and PatchCore models.
  - **Sections & Chapters:** Wrote the MR-Guided vs. MR-Only Radiology, Intensity-based candidate CT mask and Memory Bank methods.

- **Justin Verhoek**
  - **Dataset Annotation:** Performed manual annotations on the SynthRAD23 dataset.
  - **Dataset Preprocessing:** Performed dataset preprocessing including ground truth mask generation and creation of the different datasets.
  - **Knowledge Distillation Methods:** Investigated knowledge distillation-based AD approaches, implementing architectures such as RD4AD, STFPM, and MKD.
  - **Experiments with RD4AD and STFPM:** Executed the experiments from model training to custom post-processing for RD4AD and STFPM.
  - **Sections & Chapters:** Wrote the Computed Tomography (CT) Imaging, Knowledge Distillation Methods, Dataset, Ground-Truth Anomaly Mask and MR Image Preprocessing sections. Additionally, wrote the Experiments, Results, and Discussion chapters including the creation of all tables and visualizations.
- **Weijie Chen**
  - **Dataset Annotation:** Performed manual annotations on the SynthRAD23 dataset.
  - **Postprocessing:** Co-developed and implemented the post-processing pipeline with Kai Koepchen, including the slice-wise post-processing, 3d volume reconstruction and the visualization of post-processing pipeline.
  - **One-class Classification Methods:** Investigated one-class classification. AD approaches, implementing architectures such as DeepSVDD and CutPaste.
  - **Experiments with DeepSVDD and CutPaste:** Executed the experiments from model training to evaluation without applying custom post-processing.
  - **Sections & Chapters:** Wrote the Dataset overview, One-Class Classification Methods, Post-Processing (shared with Kai Koepchen), Conclusion.
- **Kai Koepchen**
  - **Dataset Annotation:** Performed manual annotations on the SynthRAD23 dataset.
  - **Postprocessing:** Co-developed and implemented the post-processing pipeline with Weijie Chen, including the 3D consistency filtering, and investigated suitable evaluation metrics and patient-level aggregation rules.
  - **Flow-Based Methods:** Investigated flow-based AD approaches, implementing architectures such as FastFlow, CFlow and CSFlow.
  - **Experiments with Fastflow and CFlow:**Executed the experiments from model training to custom post-processing for FastFlow and CFlow.
  - **Sections:** Wrote Flow-Based Methods, Anomalib Model Training, Post-Processing (shared with Weijie Chen), Evaluation, Experiment 1 and 3, Future Work and Improvements.

Appendix A

Attachments

All following tables present the complete set of metrics computed for each experiment. Detailed numerical results are placed here rather than in Chapter 7 to maintain readability of the main text, while still providing transparency. Table A.1 reports pixel-, slice-, and patient-level anomaly detection metrics for Experiment 1 before any post-processing is applied.

Table A.1: EXPERIMENT 1 RESULTS WITHOUT POST-PROCESSING. *Pixel-, slice-, and patient-level anomaly detection results for Experiment 1 without post-processing.*

Level	Model	Thr.	AUROC	Dice	Bal. Acc.	Prec.	Rec.
Pixel	DRAEM	–	0.9618	0.3504	0.7059	–	–
Pixel	Dinomaly	–	0.6269	0.0071	0.6267	–	–
Pixel	RD4AD	–	0.9539	0.0857	0.5658	–	–
Pixel	STFPM	–	0.9632	0.0907	0.8154	–	–
Pixel	CFlow	–	0.5703	0.0000	0.4999	–	–
Pixel	FastFlow	–	0.9154	0.0272	0.5107	–	–
Pixel	PatchCore	–	0.9298	0.4052	0.7257	–	–
Pixel	CFA	–	0.9692	0.0799	0.8600	–	–
Slice	DRAEM	–	0.8655	0.3860	0.5000	–	–
Slice	Dinomaly	–	0.3160	0.3860	0.5000	–	–
Slice	RD4AD	–	0.8002	0.5226	0.7091	–	–
Slice	STFPM	–	0.7380	0.3861	0.5000	–	–
Slice	CFlow	–	0.5279	0.1658	0.4750	–	–
Slice	FastFlow	–	0.6591	0.4723	0.6503	–	–
Slice	PatchCore	–	0.9821	0.7807	0.8363	–	–
Slice	CFA	–	0.7144	0.3860	0.5000	–	–
Slice	DeepSVDD	–	0.8075	0.5982	–	–	–
Slice	CutPaste	–	0.7498	0.4947	–	–	–
Patient	DRAEM	0	–	0.6364	0.5000	0.4667	1.0000
Patient	DRAEM	0.0015	–	0.8387	0.8393	0.7647	0.9286
Patient	DRAEM	0.002	–	0.8276	0.8348	0.8000	0.8571
Patient	DRAEM	0.0025	–	0.8276	0.8348	0.8000	0.8571
Patient	Dinomaly	0	–	0.6364	0.5000	0.4667	1.0000
Patient	Dinomaly	0.02	–	0.6364	0.5000	0.4667	1.0000
Patient	Dinomaly	0.025	–	0.6364	0.5000	0.4667	1.0000
Patient	Dinomaly	0.03	–	0.6364	0.5000	0.4667	1.0000

Continued on next page

Level	Model	Thr.	AUROC	Dice	Bal. Acc.	Prec.	Rec.
Patient	RD4AD	0	–	0.6364	0.5000	0.4667	1.0000
Patient	RD4AD	0.002	–	0.7273	0.7098	0.6316	0.8571
Patient	RD4AD	0.003	–	0.7586	0.7679	0.7333	0.7586
Patient	RD4AD	0.0035	–	0.7143	0.7321	0.7143	0.7143
Patient	STFPM	0	–	0.6364	0.5000	0.4667	1.0000
Patient	STFPM	0.025	–	0.7097	0.7054	0.6471	0.7857
Patient	STFPM	0.03	–	0.7143	0.7321	0.7143	0.7143
Patient	STFPM	0.035	–	0.6087	0.6875	0.7778	0.5000
Patient	CFlow	0	–	0.5714	0.5000	0.4000	1.0000
Patient	CFlow	0.001	–	0.0952	0.5263	0.5000	0.0526
Patient	CFlow	0.002	–	0.0000	0.4737	0.0000	0.0000
Patient	CFlow	0.003	–	0.0000	0.4737	0.0000	0.0000
Patient	FastFlow	0	–	0.5714	0.5000	0.4000	1.0000
Patient	FastFlow	0.001	–	0.4375	0.6281	0.5385	0.3684
Patient	FastFlow	0.002	–	0.3200	0.5746	0.6667	0.2105
Patient	FastFlow	0.003	–	0.0000	0.4737	0.0000	0.0000
Patient	PatchCore	0	–	0.6364	0.5000	0.4667	1.0000
Patient	PatchCore	0.0001	–	0.8966	0.9018	0.8667	0.9286
Patient	PatchCore	0.0005	–	0.8800	0.8929	1.0000	0.7857
Patient	PatchCore	0.001	–	0.6667	0.7500	1.0000	0.5000
Patient	CFA	0	–	0.6364	0.5000	0.4667	1.0000
Patient	CFA	0.03	–	0.6829	0.5938	0.5185	1.0000
Patient	CFA	0.04	–	0.7059	0.6786	0.6000	0.8571
Patient	CFA	0.05	–	0.4762	0.6161	0.7143	0.3571

Table A.2 reports pixel-, slice-, and patient-level AD metrics for Experiment 1 after the post-processing step has been applied.

Table A.2: EXPERIMENT 1 RESULTS. *Pixel-, slice-, and patient-level AD results for Experiment 1 after post-processing.*

Level	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Pixel	DRAEM	–	0.3911	0.7188	0.3522	0.4395
Pixel	Dinomaly	–	0.0450	0.8571	0.0231	0.7963
Pixel	RD4AD	–	0.1063	0.5659	0.0876	0.1354
Pixel	STFPM	–	0.1098	0.8318	0.0597	0.6902
Pixel	CFlow	–	0.0000	0.4999	0.0000	0.0000
Pixel	FastFlow	–	0.0303	0.5114	0.0404	0.0243
Pixel	PatchCore	–	0.4058	0.7262	0.3667	0.4544
Pixel	CFA	–	0.1990	0.8740	0.1145	0.7625
Slice	DRAEM	–	0.4417	0.6034	0.2846	0.9855
Slice	Dinomaly	–	0.3860	0.5000	0.2392	1.0000
Slice	RD4AD	–	0.5217	0.7078	0.3629	0.9277
Slice	STFPM	–	0.3862	0.5004	0.2393	1.0000
Slice	CFlow	–	0.1467	0.4710	0.1769	0.1253
Slice	FastFlow	–	0.4580	0.6355	0.3312	0.7422
Slice	PatchCore	–	0.7807	0.8363	0.8769	0.7036

Continued on next page

Level	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Slice	CFA	–	0.3873	0.5027	0.2402	1.0000
Patient	DRAEM	0	0.6364	0.5000	0.4667	1.0000
Patient	DRAEM	0.0015	0.8000	0.8036	0.7500	0.8571
Patient	DRAEM	0.002	0.8276	0.8348	0.8000	0.8571
Patient	DRAEM	0.0025	0.7407	0.7634	0.7692	0.7143
Patient	Dinomaly	0	0.6364	0.5000	0.4667	1.0000
Patient	Dinomaly	0.02	0.6316	0.5536	0.5000	0.8571
Patient	Dinomaly	0.025	0.6486	0.5848	0.5217	0.8571
Patient	Dinomaly	0.03	0.6111	0.5491	0.5000	0.7857
Patient	RD4AD	0	0.6364	0.5000	0.4667	1.0000
Patient	RD4AD	0.002	0.7500	0.7410	0.6667	0.8571
Patient	RD4AD	0.003	0.7586	0.7679	0.7333	0.7857
Patient	RD4AD	0.0035	0.7143	0.7321	0.7143	0.7143
Patient	STFPM	0	0.6364	0.5000	0.4667	1.0000
Patient	STFPM	0.025	0.7407	0.7634	0.7692	0.7143
Patient	STFPM	0.03	0.7500	0.7902	0.9000	0.6429
Patient	STFPM	0.035	0.5714	0.6830	0.8571	0.4286
Patient	CFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	CFlow	0.001	0.0952	0.5263	0.5000	0.0526
Patient	CFlow	0.002	0.1000	0.5263	1.0000	0.0526
Patient	CFlow	0.003	0.1000	0.5263	1.0000	0.0526
Patient	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	FastFlow	0.001	0.3871	0.6344	0.5000	0.3158
Patient	FastFlow	0.002	0.2963	0.5789	0.5000	0.2105
Patient	FastFlow	0.003	0.2500	0.5789	0.6000	0.1579
Patient	PatchCore	0	0.6364	0.5000	0.4667	1.0000
Patient	PatchCore	0.0001	0.8966	0.9018	0.8667	0.9286
Patient	PatchCore	0.0005	0.8800	0.8929	1.0000	0.7857
Patient	PatchCore	0.001	0.6667	0.7500	1.0000	0.5000
Patient	CFA	0	0.6364	0.5000	0.4667	1.0000
Patient	CFA	0.01	0.7429	0.7143	0.6190	0.9286
Patient	CFA	0.015	0.8276	0.8348	0.8000	0.8571
Patient	CFA	0.02	0.5714	0.6830	0.8571	0.4286

Table A.3 summarizes all metrics for Experiment 2, grouped by evaluation level and dataset format (bone colormap vs. consecutive slices), after post-processing.

Table A.3: EXPERIMENT 2 RESULTS. *Pixel-, slice-, and patient-level AD results for Experiment 2. Bone stands for the bone colormap dataset format and con stands for the consecutive adjacent slices dataset format.*

Level	Fmt.	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Pixel	con	DRAEM	–	0.1157	0.9039	0.0621	0.8386
Pixel	bone	DRAEM	–	0.0803	0.8684	0.0423	0.7798
Pixel	con	RD4AD	–	0.0882	0.5476	0.0803	0.0979

Continued on next page

Level	Fmt.	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Pixel	bone	RD4AD	–	0.1432	0.6443	0.0945	0.2955
Pixel	con	FastFlow	–	0.0030	0.4999	0.0024	0.0041
Pixel	bone	FastFlow	–	0.0005	0.4963	0.0003	0.0011
Pixel	con	PatchCore	–	0.4050	0.7169	0.3784	0.4356
Pixel	bone	PatchCore	–	0.5097	0.7850	0.4598	0.5717
Slice	con	DRAEM	–	0.3860	0.5000	0.2392	1.0000
Slice	bone	DRAEM	–	0.3860	0.5000	0.2392	1.0000
Slice	con	RD4AD	–	0.5181	0.6998	0.3734	0.8458
Slice	bone	RD4AD	–	0.4665	0.6410	0.3078	0.9639
Slice	con	FastFlow	–	0.3965	0.5466	0.2620	0.8145
Slice	bone	FastFlow	–	0.4519	0.6250	0.3093	0.8386
Slice	con	PatchCore	–	0.7750	0.8336	0.8661	0.7012
Slice	bone	PatchCore	–	0.7682	0.8245	0.8917	0.6747
Patient	con	DRAEM	0	0.6364	0.5000	0.4667	1.0000
Patient	con	DRAEM	0.025	0.6471	0.6116	0.5500	0.7857
Patient	con	DRAEM	0.03	0.7097	0.7054	0.6471	0.7857
Patient	con	DRAEM	0.035	0.5926	0.6295	0.6154	0.5714
Patient	bone	DRAEM	0	0.6364	0.5000	0.4667	1.0000
Patient	bone	DRAEM	0.03	0.7000	0.6250	0.5385	1.0000
Patient	bone	DRAEM	0.035	0.7222	0.6830	0.5909	0.9286
Patient	bone	DRAEM	0.04	0.6250	0.6071	0.5556	0.7143
Patient	con	RD4AD	0	0.6364	0.5000	0.4667	1.0000
Patient	con	RD4AD	0.0015	0.7273	0.7098	0.6316	0.8571
Patient	con	RD4AD	0.002	0.7333	0.7366	0.6875	0.7857
Patient	con	RD4AD	0.0025	0.7586	0.7679	0.7333	0.7857
Patient	bone	RD4AD	0	0.6364	0.5000	0.4667	1.0000
Patient	bone	RD4AD	0.0015	0.6316	0.5536	0.5000	0.8571
Patient	bone	RD4AD	0.002	0.6471	0.6116	0.5500	0.7857
Patient	bone	RD4AD	0.0025	0.5625	0.5402	0.5000	0.6429
Patient	con	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	con	FastFlow	0.001	0.4571	0.6754	0.5000	0.4211
Patient	con	FastFlow	0.002	0.4706	0.6929	0.5333	0.4211
Patient	con	FastFlow	0.003	0.3704	0.6316	0.6250	0.2632
Patient	bone	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	bone	FastFlow	0.001	0.1739	0.5526	0.5000	0.1053
Patient	bone	FastFlow	0.002	0.1905	0.5526	1.0000	0.1053
Patient	bone	FastFlow	0.003	0.1905	0.5526	1.0000	0.1053
Patient	con	PatchCore	0	0.6364	0.5000	0.4667	1.0000
Patient	con	PatchCore	0.0002	0.8571	0.8661	0.8571	0.8571
Patient	con	PatchCore	0.0004	0.8800	0.8929	1.0000	0.7857
Patient	con	PatchCore	0.001	0.7826	0.8214	1.0000	0.6429
Patient	bone	PatchCore	0	0.6364	0.5000	0.4667	1.0000
Patient	bone	PatchCore	0.0001	0.9333	0.9375	0.8750	1.0000
Patient	bone	PatchCore	0.0002	0.9630	0.9643	1.0000	0.9286
Patient	bone	PatchCore	0.0005	0.8333	0.8571	1.0000	0.7143

Table A.4 lists the metrics for Experiment 3, showing performance of the FastFlow model utilizing the RadImageNet backbone across the replicated, bone colormap, and consecutive-slice input formats at pixel-, slice-, and patient-level.

Table A.4: EXPERIMENT 3 RESULTS. *Pixel-, slice-, and patient-level AD results for Experiment 3. Rep stands for the baseline replicated-channel dataset format, bone stands for the bone colormap dataset format and con stands for the consecutive adjacent slices dataset format.*

Level	Fmt.	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Pixel	rep	FastFlow	0	0.0327	0.5233	0.0239	0.0518
Pixel	con	FastFlow	0.001	0.0796	0.6407	0.0460	0.2964
Pixel	bone	FastFlow	0.002	0.0457	0.5447	0.0299	0.0970
Slice	rep	FastFlow	0	0.4500	0.6251	0.3169	0.7759
Slice	con	FastFlow	0.001	0.4319	0.5921	0.2821	0.9205
Slice	bone	FastFlow	0.002	0.4404	0.6130	0.3070	0.7783
Patient	rep	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	rep	FastFlow	0.001	0.3871	0.6344	0.5000	0.3158
Patient	rep	FastFlow	0.002	0.2963	0.5789	0.5000	0.2105
Patient	rep	FastFlow	0.003	0.2500	0.5789	0.6000	0.1579
Patient	con	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	con	FastFlow	0.001	0.4706	0.6929	0.5333	0.4211
Patient	con	FastFlow	0.002	0.4138	0.6579	0.6000	0.3158
Patient	con	FastFlow	0.003	0.4444	0.6579	0.7500	0.3158
Patient	bone	FastFlow	0	0.7037	0.5000	0.5429	1.0000
Patient	bone	FastFlow	0.001	0.2400	0.5789	0.5000	0.1579
Patient	bone	FastFlow	0.002	0.1818	0.5526	0.6667	0.1053
Patient	bone	FastFlow	0.003	0.1905	0.5526	1.0000	0.1053

Table A.5 reports the complete pixel-, slice-, and patient-level metrics for Experiment 4 using the multi-center dataset, for all evaluated models.

Table A.5: EXPERIMENT 4 RESULTS. *Pixel-, slice-, and patient-level AD results for Experiment 4. Rep stands for the replicated-channel dataset format and bone stands for the bone colormap dataset format.*

Level	Fmt.	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Pixel	rep	DRAEM	–	0.1832	0.7043	0.1176	0.4138
Pixel	rep	RD4AD	–	0.0318	0.5139	0.0351	0.0291
Pixel	rep	FastFlow	–	0.0001	0.4981	0.0006	0.0021
Pixel	bone	PatchCore	–	0.4856	0.7899	0.4171	0.5811
Slice	rep	DRAEM	–	0.2925	0.5213	0.1713	1.0000
Slice	rep	RD4AD	–	0.4034	0.6736	0.2851	0.6882
Slice	rep	FastFlow	–	0.2868	0.5143	0.1695	0.9301
Slice	bone	PatchCore	–	0.7099	0.8202	0.7301	0.6909
Patient	rep	DRAEM	0	0.5614	0.5000	0.3902	1.0000
Patient	rep	DRAEM	0.0015	0.5818	0.5400	0.4103	1.0000
Patient	rep	DRAEM	0.002	0.5926	0.5600	0.4211	1.0000
Patient	rep	DRAEM	0.0025	0.5600	0.5375	0.4118	0.8750

Continued on next page

Level	Fmt.	Model	Thr.	Dice	Bal. Acc.	Prec.	Rec.
Patient	rep	RD4AD	0	0.5614	0.5000	0.3903	1.0000
Patient	rep	RD4AD	0.001	0.5789	0.6238	0.5000	0.6875
Patient	rep	RD4AD	0.0015	0.6452	0.7125	0.6667	0.6250
Patient	rep	RD4AD	0.002	0.5714	0.6700	0.6667	0.5000
Patient	rep	FastFlow	0	0.5614	0.5	0.3901	1.0000
Patient	rep	FastFlow	0.001	0.58824	0.5686	0.4286	0.9375
Patient	rep	FastFlow	0.002	0.5532	0.54625	0.4194	0.8125
Patient	rep	FastFlow	0.003	0.5116	0.5238	0.4074	0.6875
Patient	bone	PatchCore	0	0.5614	0.5000	0.3902	1.0000
Patient	bone	PatchCore	0.0002	0.7568	0.7975	0.6667	0.8750
Patient	bone	PatchCore	0.0004	0.7742	0.8150	0.8000	0.7500
Patient	bone	PatchCore	0.0005	0.7586	0.8038	0.8462	0.6875

List of Figures

List of Tables

Bibliography

Akcay, S., Ameln, D., Vaidya, A., Lakshmanan, B., Ahuja, N., and Genc, U. (2022). Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1706–1710. IEEE.

Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., and Li, X. (2024). Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053.

Barusco, M., Borsatti, F., Beda, N., Dalle Pezze, D., and Susto, G. A. (2025). Towards continual visual anomaly detection in the medical domain. In *Proceedings of the International Workshop on Personalized Incremental Learning in Medicine*, pages 18–24.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192.

Cao, T., Huang, C.-W., Hui, D. Y.-T., and Cohen, J. P. (2020). A benchmark of medical out of distribution detection.

Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey.

Chandarana, H., Wang, H., Tijssen, R., and Das, I. J. (2018). Emerging role of mri in radiation therapy. *Journal of Magnetic Resonance Imaging*, 48(6):1468–1478.

Chen, C. A., Chen, W., Goodman, S. B., Hargreaves, B. A., Koch, K. M., Lu, W., Brau, A. C., Draper, C. E., Delp, S. L., and Gold, G. E. (2011). New mr imaging methods for metallic implants in the knee: artifact correction and clinical impact. *Journal of Magnetic Resonance Imaging*, 33(5):1121–1127.

Chin, S., Eccles, C. L., McWilliam, A., Chuter, R., Walker, E., Whitehurst, P., Berresford, J., Van Herk, M., Hoskin, P. J., and Choudhury, A. (2020). Magnetic resonance-guided radiation therapy: A review. *Journal of Medical Imaging and Radiation Oncology*, 64(1):163–177.

Chourak, H., Barateau, A., Nunes, J.-C., Greer, P. B., Tahri, S., Lafond, C., de Crevoisier, R., Dowling, J., and Acosta, O. (2022a). Local quality assessment of patient specific synthetic-ct via voxel-wise analysis. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7.

Chourak, H., Barateau, A., Tahri, S., Cadin, C., Lafond, C., Nunes, J.-C., Boue-Rafle, A., Perazzi, M., Greer, P. B., Dowling, J., et al. (2022b). Quality assurance for mri-only radiation therapy: A voxel-wise population-based methodology for image and dose assessment of synthetic ct generation methods. *Frontiers in oncology*, 12:968689.

Dal Bello, R., Lapaeva, M., La Greca Saint-Esteven, A., Wallimann, P., Günther, M., Konukoglu, E., Andratschke, N., Guckenberger, M., and Tanadini-Lang, S. (2023). Patient-specific quality assurance strategies for synthetic computed tomography in magnetic resonance-only radiotherapy of the abdomen. *Physics and Imaging in Radiation Oncology*, 27:100464.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer.

Deng, H. and Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. In *International Conference on Learning Representations (ICLR)*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Fusella, M., Alvarez Andres, E., Villegas, F., Milan, L., Janssen, T., Dal Bello, R., Garibaldi, C., Placidi, L., and Cusumano, D. (2024). Results of 2023 survey on the use of synthetic computed tomography for magnetic resonance imaging-only radiotherapy: Current status and future steps. *Physics and Imaging in Radiation Oncology*, 32:100652.

Goldman, L. W. (2007). Principles of ct and ct technology. *Journal of nuclear medicine technology*, 35(3):115–128.

Gudovskiy, D., Ishizaka, S., and Kozuka, K. (2022). Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. pages 98–107.

Guo, B., Lu, D., Szumel, G., Gui, R., Wang, T., Konz, N., and Mazurowski, M. A. (2024). The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *arXiv preprint arXiv:2409.04368*.

Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., and Liao, H. (2025). Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20405–20415.

Hall, W. A., Paulson, E. S., van der Heide, U. A., Fuller, C. D., Raaymakers, B., Lagend, J. J., Li, X. A., Jaffray, D. A., Dawson, L. A., Erickson, B., Verheij, M., Harrington, K. J., Sahgal, A., Lee, P., Parikh, P. J., Bassetti, M. F., Robinson, C. G., Minsky, B. D., Choudhury, A., Tersteeg, R. J., and Schultz, C. J. (2019). The transformation of radiation oncology using real-time magnetic resonance guidance: A review. *European Journal of Cancer*, 122:42–52.

Hargreaves, B. A., Worters, P. W., Pauly, K. B., Pauly, J. M., Koch, K. M., and Gold, G. E. (2011). Metal-induced artifacts in mri. *American Journal of Roentgenology*, 197(3):547–555. PMID: 21862795.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heckler, L., König, R., and Bergmann, P. (2023). Exploring the importance of pretrained feature extractors for unsupervised anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2926.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Höfler, D., Grigo, J., Siavosch, H., Saake, M., Schmidt, M. A., Weissmann, T., Schubert, P., Voigt, R., Lettmaier, S., Semrau, S., et al. (2025). Mri distortion correction is associated with improved local control in stereotactic radiotherapy for brain metastases. *Scientific Reports*, 15(1):9077.

Huijben, E. M., Terpstra, M. L., Galapon, A. J., Pai, S., Thummerer, A., Koopmans, P., Afonso, M., van Eijnatten, M., Gurney-Champion, O., Chen, Z., Zhang, Y., Zheng, K., Li, C., Pang, H., Ye, C., Wang, R., Song, T., Fan, F., Qiu, J., Huang, Y., Ha, J., Sung Park, J., Alain-Beaudoin, A., Bériault, S., Yu, P., Guo, H., Huang, Z., Li, G., Zhang, X., Fan, Y., Liu, H., Xin, B., Nicolson, A., Zhong, L., Deng, Z., Müller-Franzes, G., Khader, F., Li, X., Zhang, Y., Hémon, C., Boussot, V., Zhang, Z., Wang, L., Bai, L., Wang, S., Mus, D., Kooiman, B., Sargeant, C. A., Henderson, E. G., Kondo, S., Kasai, S., Karimzadeh, R., Ibragimov, B., Helfer, T., Dafflon, J., Chen, Z., Wang, E., Perko, Z., and Maspero, M. (2024). Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report. *Medical Image Analysis*, 97:103276.

Hémon, C., Cubero, L., Boussot, V., Martin, R.-A., Texier, B., Castelli, J., de Crevoisier, R., Barateau, A., Lafond, C., and Nunes, J.-C. (2025). Modeling dose uncertainty in cone-beam computed tomography: Predictive approach for deep learning-based synthetic computed tomography generation. *Physics and Imaging in Radiation Oncology*, 33:100704.

Johnstone, E., Wyatt, J. J., Henry, A. M., Short, S. C., Sebag-Montefiore, D., Murray, L., Kelly, C. G., McCallum, H. M., and Speight, R. (2018). Systematic review of synthetic computed tomography generation methodologies for use in magnetic resonance imaging–only radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 100(1):199–217.

Jonsson, J., Nyholm, T., and Söderkvist, K. (2019). The rational for mr-only treatment planning for external radiotherapy. *Clinical and Translational Radiation Oncology*, 18.

Jung, H. (2021). Basic physical principles and clinical applications of computed tomography. *Progress in Medical Physics*, 32(1):1–17.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Kim, B., Kwon, K., Oh, C., and Park, H. (2021). Unsupervised anomaly detection in mr images using multicontrast information. *Medical Physics*, 48(11):7346–7359.

Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. (2021). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

Kocak, B., Klontzas, M. E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., Bressem, K., Ugga, L., Mercaldo, N., Díaz, O., et al. (2025). Evaluation metrics in medical imaging ai: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, page 100030.

La Greca Saint-Esteven, A., Dal Bello, R., Lapaeva, M., Fankhauser, L., Pouymayou, B., Konukoglu, E., Andratschke, N., Balermpas, P., Guckenberger, M., and Tanadini-Lang, S. (2023). Synthetic computed tomography for low-field magnetic resonance-only radiotherapy in head-and-neck cancer using residual vision transformers. *Physics and Imaging in Radiation Oncology*, 27:100471.

Lapaeva, M. (2025). *Proof-of-Concept for Automating Patient-Specific Quality Assurance in MR-Only Radiotherapy Using Deep Learning-Based Out-of-Distribution Detection*, page 45–45. CERN.

Lapaeva, M., La Greca Saint-Esteven, A., Wallimann, P., Andratschke, N., Guckenberger, M., Günther, M., Tanadini-Lang, S., and Dal Bello, R. (2025). A comprehensive comparative study of generative adversarial network architectures for synthetic computed tomography generation in the abdomen. *Medical Physics*, 52(8):e18038.

Lapaeva, M., La Greca Saint-Esteven, A., Wallimann, P., Günther, M., Konukoglu, E., Andratschke, N., Guckenberger, M., Tanadini-Lang, S., and Dal Bello, R. (2022). Synthetic computed tomography for low-field magnetic resonance-guided radiotherapy in the abdomen. *Physics and Imaging in Radiation Oncology*, 24:173–179.

Law, M. W., Tse, M. Y., Ho, L. C. C., Cheng, W., Chan, N., Yeung, W.-L., Tse, E., Tsang, S., Lam, K., and Yu, K. (2024). A study of bayesian deep network uncertainty and its application to synthetic ct generation for mr-only radiotherapy treatment planning. *Medical Physics*, 51(2):1244–1262.

Lee, S., Lee, S., and Song, B. C. (2022). Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021a). Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674.

Li, N., Jiang, K., Ma, Z., Wei, X., Hong, X., and Gong, Y. (2021b). Anomaly detection via self-organizing map. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 974–978. IEEE.

Li, X., Bellotti, R., Meier, G., Bachtiary, B., Weber, D., Lomax, A., Buhmann, J., and Zhang, Y. (2024). Uncertainty-aware mr-based ct synthesis for robust proton therapy planning of brain tumour. *Radiotherapy and Oncology*, 191:110056.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Zhang, Y., Yang, Y., Li, H., Zhao, Q., Xu, Z., et al. (2022). Radimagenet: An open radiologic deep learning research dataset for transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315.

Messaoudi, H., Belaid, A., Ben Salem, D., and Conze, P.-H. (2023). Cross-dimensional transfer learning in medical image segmentation with deep learning. *Medical Image Analysis*, 90:102868.

Michael, G. (2001). X-ray computed tomography. *Physics Education*, 36(6):442.

Nella, F., Tanadini-Lang, S., and Dal Bello, R. (2025). Clinical implementation of patient-specific quality assurance for synthetic computed tomography. *Physics and Imaging in Radiation Oncology*, 34:100764.

Nyholm, T., Nyberg, M., Karlsson, M. G., and Karlsson, M. (2009). Systematisation of spatial uncertainties for comparison between a mr and a ct-based radiotherapy workflow for prostate treatments. *Radiation oncology*, 4(1):54.

Ocanto, A., Torres, L., Montijano, M., Rincón, D., Fernández, C., Sevilla, B., Gonsalves, D., Teja, M., Guijarro, M., Glaría, L., Hernánz, R., Zafra-Martin, J., Sanmamed, N., Kishan, A., Alongi, F., Moghanaki, D., Nagar, H., and Couñago, F. (2024). Mr-linac, a new partner in radiation oncology: Current landscape. *Cancers*, 16(2).

Ottoni, M., Kasperczuk, A., and Tavora, L. M. (2025). Machine learning in mri brain imaging: A review of methods, challenges, and future directions. *Diagnostics*, 15(21):2692.

Paradis, E., Cao, Y., Lawrence, T. S., Tsien, C., Feng, M., Vineberg, K., and Balter, J. M. (2015). Assessing the dosimetric accuracy of magnetic resonance-generated synthetic ct images for focal brain vmat radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 93(5):1154–1161.

Parchur, A. K., Zarenia, M., Gage, C., Paulson, E. S., and Ahunbay, E. (2025). Automated hallucination detection for synthetic ct images used in mr-only radiotherapy workflows. *Physics in Medicine & Biology*, 70(5):05NT01.

Pereira, G. C., Traughber, M., and Muzic Jr, R. F. (2014). The role of imaging in radiation therapy planning: past, present, and future. *BioMed research international*, 2014(1):231090.

Perlin, K. (1985). An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296.

Placidi, L., Cusumano, D., Boldrini, L., Votta, C., Pollutri, V., Antonelli, M. V., Chiloiro, G., Romano, A., De Luca, V., Catucci, F., Indovina, L., and Valentini, V. (2020). Quantitative analysis of mri-guided radiotherapy treatment process time for tumor real-time gating efficiency. *Journal of Applied Clinical Medical Physics*, 21(11):70–79.

Prerau, M. J. and Eskin, E. (2000). Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm. *Undergraduate Thesis, Columbia University: December*.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. (2015). Fitnets: Hints for thin deep nets.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. (2022). Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328.

Rudd, E. M., Jain, L. P., Scheirer, W. J., and Boult, T. E. (2018). The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):762–768.

Rudolph, M., Wandt, B., and Rosenhahn, B. (2021). Same same but differnet: Semi-supervised defect detection with normalizing flows. pages 1907–1916.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., and Rabiee, H. R. (2021). Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (1999). Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 582–588.

Soille, P. (2003). *Morphological Image Analysis: Principles and Applications*. Springer, Berlin, Heidelberg, 2 edition.

Spadea, M. F., Maspero, M., Zaffino, P., and Seco, J. (2021). Deep learning based synthetic-ct generation in radiotherapy and pet: A review. *Medical Physics*, 48(11):6537–6566.

Taha, A. A. and Hanbury, A. (2015). Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29.

Tax, D. and Duin, R. (2004). Support vector data description. *Machine Learning*, 54:45–66.

Teng, Y., Li, H., Cai, F., Shao, M., and Xia, S. (2022). Unsupervised visual defect detection with score-based generative model.

Thummerer, A., van der Bijl, E., Galapon, A., Verhoeff, J. J. C., Langendijk, J. A., Both, S., van den Berg, C. N. A. T., and Maspero, M. (2023). Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical Physics*, 50(7):4664–4674.

Thummerer, A., van der Bijl, E., Galapon, A. J., Kamp, F., Savenije, M., Muijs, C., Aluwini, S., Steenbakkers, R. J. H. M., Beuel, S., Intven, M. P. W., Langendijk, J. A., Both, S., Corradini, S., Rogowski, V., Terpstra, M., Wahl, N., Kurz, C., Landry, G., and Maspero, M. (2025). Synthrad2025 grand challenge dataset: generating synthetic cts for radiotherapy.

van Harten, L. D., Wolterink, J. M., Verhoeff, J. J., and Išgum, I. (2020). Automatic online quality control of synthetic cts. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 399–405. SPIE.

Vasiliuk, A., Frolova, D., Belyaev, M., and Shirokikh, B. (2023). Limitations of out-of-distribution detection in 3d medical image segmentation. *Journal of Imaging*, 9(9):191.

Walker, A., Chlap, P., Causer, T., Mahmood, F., Buckley, J., and Holloway, L. (2022). Development of a vendor neutral mri distortion quality assurance workflow. *Journal of Applied Clinical Medical Physics*, 23(10):e13735.

Wang, C., Chao, M., Lee, L., and Xing, L. (2008). Mri-based treatment planning with electron density information mapped from ct images: A preliminary study. *Technology in Cancer Research & Treatment*, 7(5):341–347. PMID: 18783283.

Wang, G., Han, S., Ding, E., and Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*.

Wang, H., Balter, J., and Cao, Y. (2013). Patient-induced susceptibility effect on geometric distortion of clinical brain mri for radiation treatment planning on a 3t scanner. *Physics in Medicine & and Biology*, 58(3):465.

Wang, S., Xu, X., Zhang, L.-y., Du, H., Chen, Y., and Mei, W. (2025). Enhancing patient-level diagnosis on mr images: a multi-instance learning framework with contributive feature mining. *Measurement Science and Technology*, 36:045701.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612.

Weishaupt, D., Köchli, V., Marincek, B., and Kim, E. E. (2007). How does mri work? an introduction to the physics and function of magnetic resonance imaging. *Journal of Nuclear Medicine*, 48(11):1910–1910.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., and Le, X. (2022a). A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584.

You, Z., Yang, K., Luo, W., Cui, L., Zheng, Y., and Le, X. (2022b). Adtr: Anomaly detection transformer with feature reconstruction. In *International Conference on Neural Information Processing*, pages 298–310. Springer.

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. (2022). Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. pages 9930–9940.

Zaffino, P., Raggio, C. B., Thummerer, A., Marmitt, G. G., Langendijk, J. A., Procopio, A., Cosentino, C., Seco, J., Knopf, A. C., Both, S., and Spadea, M. F. (2024). Toward closing the loop in image-to-image conversion in radiotherapy: A quality control tool to predict synthetic computed tomography hounsfield unit accuracy. *Journal of Imaging*, 10(12).

Zavrtanik, V., Kristan, M., and Skočaj, D. (2021). Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339.

Zavrtanik, V., Kristan, M., and Skočaj, D. (2020). Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M.-H., and Tao, D. (2024). Exploring plain vit reconstruction for multi-class unsupervised anomaly detection.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674.