Latent Diffusion Models for Weakly Supervised Kidney Anomaly Detection

Master's Thesis Data Science

Jen Dusseljee
s1003489

December 11, 2025

Daily supervisor:
drs. Sarah de Boer

First supervisor/assessor:
dr. Alessa Hering

Second assessor:
dr. Colin Jacobs

Radboud Universiteit Radboudumc

Abstract

Automatic detection of kidney cancer on computed tomography (CT) scans can enhance diagnostic workflows, particularly as kidney tumors are frequently discovered incidentally during unrelated imaging. While artificial intelligence algorithms show promise in supporting radiologists, most methods rely on supervised learning and are constrained by the availability and diversity of annotated datasets. Weakly supervised anomaly detection offers an alternative by learning from large unannotated datasets, with recent work demonstrating the use of diffusion models for anomaly detection through reconstruction of healthy versions of pathological images and computation of voxel-wise differences.

To the best of our knowledge, this thesis presents the first fully 3D anomaly detection pipeline for abdominal CT based on latent diffusion models. Our approach combines VQ-GAN autoencoders with DDPM and DDIM sampling to perform reconstruction-based anomaly detection on full 3D kidney volumes. The pipeline uses TotalSegmentator to select regions of interest around each kidney and is trained using pseudo-labels automatically derived from radiology reports, eliminating the need for manual annotations. We benchmark our methods against state-of-the-art supervised models (nnU-Net and nnDetection) across two datasets, conducting comprehensive evaluation including size-stratified analysis and component-wise assessment.

Our results show that while diffusion-based methods achieve some detection capability, they significantly underperform supervised baselines, with DSC scores of $0.07-0.12$ compared to 0.68 for nnU-Net, and F1 scores of $0.02-0.03$ compared to $0.63-0.69$ for nnU-Net and nnDetection. Contrary to existing literature, our DDPM-based method outperformed our DDIM-based method in our evaluation, which we attribute to the poor performance of our classifier (AUC = 0.56). Our qualitative analysis demonstrates the promise of diffusion-based reconstruction while identifying key issues.

While not achieving competitive performance or clinical viability, this work provides valuable insights into the limitations of current approaches and establishes a foundation for future improvements in diffusion-based medical anomaly detection.

Acknowledgements

I would like to express my sincere gratitude to everyone at the Diagnostic Image Analysis Group at Radboudumc for inspiring my pursuit of a career in AI for medical imaging and for creating such a stimulating research environment. I am particularly grateful to my supervisors, Sarah de Boer and Alessa Hering, who provided not only invaluable support and guidance throughout this thesis but also meaningful mentorship and career advice.

I also wish to thank the ELLIS Unit Nijmegen for the opportunity to participate in their Excellence Fellowship program, which enabled me to exchange ideas and share my research findings with fellow motivated students.

Finally, I would like to acknowledge the COMFORT project team for their collaboration and for providing me with valuable insights into current research on AI for urologic care.

Contents

Chapter 1

Introduction

Kidney cancer has a yearly incidence rate of approximately 400,000 new cases worldwide [14]. Because patients often remain asymptomatic until advanced stages, kidney tumors are frequently detected incidentally during imaging performed for unrelated medical reasons [11]. Early and automatic detection of kidney tumors on computed tomography (CT) scans could therefore provide substantial clinical value by supporting radiologists, particularly in identifying small lesions that might otherwise be overlooked.

Current artificial intelligence approaches for kidney lesion detection rely primarily on supervised learning, requiring extensive manual annotations by medical experts [17, 19, 53]. While supervised methods like nnU-Net achieve strong performance [32, 10], they face significant limitations: annotation requires substantial time investments and medical expertise, is subject to inter-rater variability [33, 43], and may generalize poorly to rare or unseen abnormalities due to limited training data diversity [22].

Weakly supervised anomaly detection offers a promising alternative by learning from large unannotated datasets. These methods identify lesions as deviations from healthy anatomy using reconstruction-based approaches with generative models. Recent advances in denoising diffusion models, both probabilistic ones (DDPM) and implicit ones (DDIM), have shown superior performance compared to traditional approaches like VAEs [13, 36] and GANs [23, 54] in both medical image generation [34] and anomaly detection [67, 68].

However, current diffusion-based anomaly detection methods face three critical limitations that prevent clinical adoption. First, existing approaches operate predominantly on individual 2D slices, while state-of-the-art supervised methods process full 3D volumes to capture complex anatomical relationships. Second, existing research has focused primarily on brain MRI, a relatively homogeneous imaging domain, with limited exploration of challenging anatomical regions like the abdomen where tissue heterogeneity and organ complexity pose greater reconstruction challenges. Finally, the field

lacks benchmarking against supervised state-of-the-art methods, making it difficult to assess whether these diffusion-based approaches can compete with established methods.

In this thesis, we address these limitations and make the following contributions:

1. We propose the first 3D latent diffusion pipeline for kidney lesion detection. Our pipeline combines TotalSegmentator to automatically identify kidney regions of interest, 3D latent diffusion models trained on healthy kidney regions, and classifier guidance during reconstruction to identify anomalous regions. We implement and compare both DDPM and DDIM sampling methods, and train the system using only pseudo-labels derived from routine clinical CT scans with radiological reports, eliminating the need for manual annotations.

2. We perform a benchmark comparison against supervised segmentation and detection models, a gap often overlooked in literature on diffusion-based anomaly detection.

3. We perform an elaborate analysis of the sub-components of the pipeline, highlighting the flaws in the method and aiding future research into this topic.

1.1 Research Questions

To systematically evaluate our proposed approach, we formulate the following research questions grouped into three categories:

1. Latent diffusion-based kidney anomaly detection performance.

- **RQ1a** Which of the two sampling methods (DDPM, DDIM) reaches the highest segmentation performance on the test sets?
- **RQ1b** Which of the two sampling methods (DDPM, DDIM) reaches the highest detection performance on the test sets?
- **RQ1c** How does the performance of the models vary with different lesion sizes, based on the TNM staging system for renal cancer?

2. Comparison against supervised baselines.

- **RQ2a** Does a dedicated detection model, like nnDetection, outperform the existing nnU-Net baseline on detection performance?
- **RQ2b** Can weakly supervised, diffusion-based methods perform on par with supervised segmentation methods, like nnU-Net?
- **RQ2c** Can weakly supervised, diffusion-based methods perform on par with supervised detection methods, like nnDetection and nnU-Net?

3. Individual subcomponent performance.

RQ3a What is the generative performance of our latent diffusion model?

RQ3b How well does the classifier perform at distinguishing healthy from unhealthy kidneys in the noised latent space?

RQ3c How well does TotalSegmentator detect kidneys in the test datasets?

1.2 Thesis Structure

The structure of this thesis is as follows: Chapter 2 provides the necessary background on CT imaging, kidney anatomy, and the core generative models used in our approach. Chapter 3 reviews related work in both supervised and unsupervised medical detection tasks, highlighting relevant work on kidney lesion detection and diffusion based anomaly detection in medical imaging. Chapter 4 details our proposed 3D latent diffusion pipeline. This chapter also describes the datasets we use and how we generate pseudo-labels for our weakly-supervised method. Chapter 5 outlines our experimental methodology, including hyperparameter optimization, the evaluation and benchmarking of our complete pipeline, evaluation on different lesion sizes, and the evaluation of the sub-components that make up the pipeline. Chapter 6 presents and interprets the results for all the experiments described in Chapter 5. Finally, in Chapter 7, we answer our research questions and discuss the implications of our results, limitations and future directions, while Chapter 8 summarizes our key findings.

Chapter 2

Background

2.1 Clinical background

This section serves to provide the reader with the clinical background knowledge required to understand the rest of the thesis. First, the basics of CT are covered, followed by an introduction to kidney anatomy, function, and lesions.

2.1.1 Computed Tomography

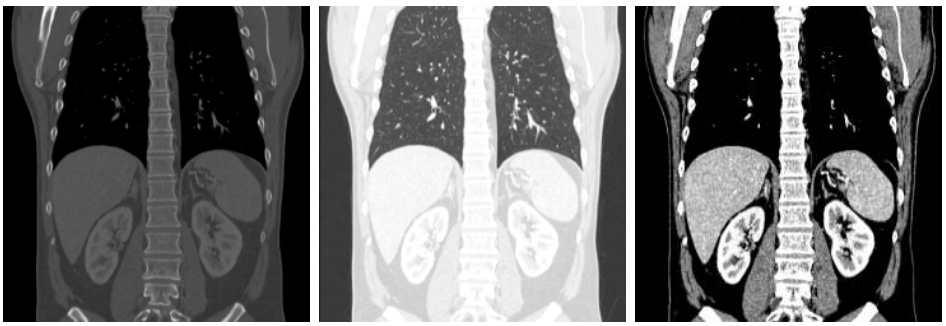
Modern medicine relies on several imaging methods to visualize the interior of a patient's body without the need for invasive surgery. Common medical imaging modalities include MRI (Magnetic Resonance Imaging), ultrasound, X-ray, and CT (Computed Tomography). This thesis focuses on CT, a medical imaging technique that uses multiple X-rays to compose a volumetric image of the body.

(a) Axial (b) Coronal (c) Sagittal

Figure 2.1: **CT image of the same anatomical region displayed in the three orthogonal planes.** a) axial plane (bottom-to-top), b) coronal plane (front-to-back), and c) sagittal plane (right-to-left).

During a CT scan, a detector ring acquires multiple X-ray images from different orientations, which are reconstructed into a single cross-sectional slice. By advancing the patient through the detector ring, multiple slices are acquired, which are subsequently composed into a single 3D image that can be examined from different planes (see Figure 2.1).

The intensity of each voxel (3D pixel) is represented in Hounsfield Units (HU), which quantify the radiodensity of the tissue relative to water (0 HU) and air (-1000 HU). Denser tissues such as bone exhibit positive HU values (typically 1000+ HU), while less dense tissues such as fat exhibit negative values (around -100 HU). When viewing CT scans, a window with level L and width W is selected; only HU values within this range (L$-$W/2 to L+W/2) are displayed. The selection of different windows improves the visibility of details for specific tissue types (see Figure 2.2).



(a) Bone window (W:1800 L:400) (b) Lung window (W:1500 L:-600) (c) Soft tissue window (W:250 L:50)

Figure 2.2: **CT image of the same anatomical region displayed in the coronal plane, visualized with different window levels and widths.** The window selection allows enhanced detail visualization for different tissue types.

Several acquisition and processing parameters influence the visual quality of CT images. The spatial resolution and slice thickness are fundamental factors that determine voxel dimensions. These parameters are constrained by the CT scanner specifications, with older systems requiring larger voxels to maintain adequate signal-to-noise ratios, while modern scanners can achieve higher spatial resolution and thinner slice thickness while preserving image quality. However, the highest available resolution is not always optimal due to increased radiation exposure to the patient and elevated noise levels. The reconstruction kernel represents another critical parameter in the conversion of raw CT data to the final image. Kernel selection involves a trade-off between image characteristics. Sharp or edge-enhancing kernels excel at preserving fine anatomical details and boundary definition but introduce texture noise in homogeneous tissue regions, whereas smooth kernels

optimize the visualization of large anatomical structures and reduce noise at the cost of spatial resolution and edge sharpness.

To improve the visibility of certain anatomical structures, a contrast agent can be administered to the patient. This contrast agent exhibits high radiodensity in CT imaging, thereby enhancing the visualization of specific anatomical structures. The time interval between contrast administration and CT scan acquisition determines which anatomical structures are enhanced. This timing can be broadly categorized into the following ranges, referred to as contrast phases [55]:

- **Early arterial phase:** 15-20 seconds after injection.
- **Late arterial phase:** 35-40 seconds after injection.
- **Hepatic or late portal phase:** 70-80 seconds after injection.
- **Nephrogenic phase:** 100 seconds after injection.
- **Delayed phase:** 6-10 minutes after injection.

2.1.2 Kidneys and Kidney Lesions

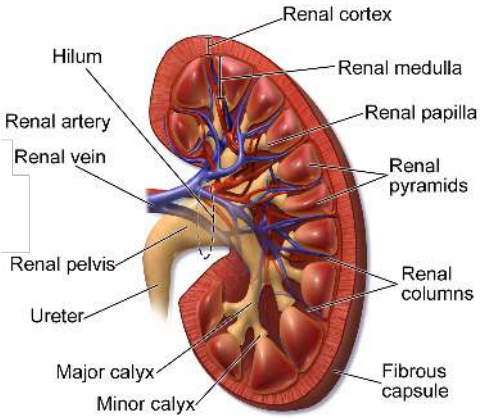


Figure 2.3: **Cross-sectional anatomy of the kidney showing major structural components.** The main regions of interest are the renal cortex, renal medulla, renal pelvis, and the renal artery and vein connecting at the hilum. Adapted from [9].

Both kidneys and ureters are clearly visible on most CT scans. Contrast enhancement is frequently employed to distinguish between the medulla and cortex.

Kidney lesions are defined as any areas of the kidney that are visually distinct from normal, healthy kidney tissue. These may manifest as cysts, fluid-filled sacs that are typically benign, or tumors, solid masses that are

potentially malignant [50]. Tumors can be further classified as either benign (such as adenomas, oncocytomas, and angiomyolipomas [42]) or malignant. Differential diagnosis is performed through evaluation of the lesion using several imaging modalities, including CT with different contrast phases, MRI, and PET scans. Malignancy can also be confirmed through histopathological analysis of tissue samples acquired via biopsy. Malignant tumors that originate within the kidney are classified as renal cell carcinomas (RCC), whereas tumors that have metastasized from other primary sites are termed metastases.

TNM staging for kidney cancer

The TNM (Tumor, Node, Metastasis) staging system is used to determine the severity of a cancer [60]. The TNM stage is made up of three components, according to criteria specific to the organ in which the cancer originated. In this subsection, we cover the TNM staging system for RCC specifically.

First, the T stage describes the size of the tumor:

T1 The tumor is under 7cm in diameter and falls completely within the kidney.

- **T1a** The tumor is 4cm in diameter or smaller.
- **T1b** The tumor is between 4cm and 7cm in diameter.

T2 The tumor is larger than 7cm in diameter, but still falls completely within the kidney.

- **T2a** The tumor is between 7cm and 10cm in diameter.
- **T2b** The tumor is over 10cm in diameter.

T3 The tumor has grown outside of the kidney, including the renal vein or the vein that takes blood back to the heart (vena cava). The tumor has not grown into the adrenal gland or outside the tissue encapsulating the kidney (fascia).

- **T3a** The tumor has grown into nearby tissue or the renal vein.
- **T3b** The tumor has grown into the vena cava.
- **T3c** The tumor has grown into the vena cava beyond to above the diaphragm or has grown into the wall of the vena cava.

T4 The tumor has grown into the tissue beyond the fascia, including potentially the adrenal gland.

Second, the N stage describes lymph node involvement:

N0 There are no cancer cells present in the lymph nodes close to the kidney.

N1 Cancer cells are present in the lymph nodes close to the kidney.

Lastly, the M stage describes the presence of metastases:

M0 The cancer has not metastasized to other organs.

M1 The cancer has metastasized to other organs.

2.2 Generative models

The objective of generative models is to enable the generation of new samples that possess the same properties as samples from an existing dataset. This process can be conceptualized as sampling from the distribution $p(x)$ of the existing dataset. In practice, direct sampling from $p(x)$ is intractable, since the exact data distribution is unknown. Therefore, instead of sampling directly from $p(x)$, a generative model p_θ is learned that takes a sample from the Gaussian distribution $z \sim p(z) = \mathcal{N}(0, I)$ and models the distribution $p_\theta(x|z)$.

In this section, the mathematical background of the three most important generative models necessary to understand the rest of the thesis is presented. The following subsections explain Auto-Encoders, Adversarial Networks, and Diffusion Models individually and then describe how these techniques combine into the architecture used in this thesis, called Latent Diffusion.

2.2.1 Auto-Encoders

A traditional Auto-Encoder (AE) consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ network. The encoder network transforms input data x to a feature vector $z = \mathcal{E}(x)$ (also called the bottleneck) with a smaller size. The decoder network in turn transforms z into output data $x' = \mathcal{D}(z)$ with the same size as x. During training, the model is taught to minimize the difference between x and x', typically using a loss function, e.g. the Mean-Squared Error (MSE) loss: $\mathcal{L}(x, x') = ||x - x'||^2$. This full process is depicted in Figure 2.4 The resulting model can then be used for dimensionality reduction or feature learning. Since the only training objective for an AE is the minimization of the reconstruction loss, no constraints are placed on the distribution of the latent variables z. This results in the latent space of an AE having no reliable structure, and sampling from it is not guaranteed to produce meaningful output after decoding [36]. These problems are mitigated by Variational Auto-Encoders, as introduced in the following section.

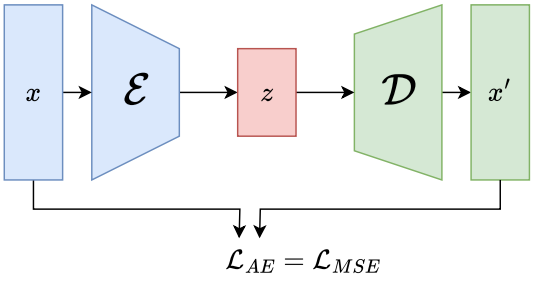


Figure 2.4: **The auto-encoder architecture.** The encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are trained jointly to learn a conversion between image space (x and x') and a latent space (z) by minimizing the reconstruction loss $\mathcal{L}_{MSE}$.

Variational Auto-Encoders

Variational Auto-Encoders [36] (VAEs) are a member of the auto-encoder family that uses probabilistic modeling. Where a traditional Auto-Encoder maps a given input x to a specific vector z in latent space, a Variational Auto-Encoder maps x to a latent distribution $q(z|x)$, given by the multivariate gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$ where μ is the mean vector and σ^2 is the variance vector. The latent vector z is then a sample from this latent distribution $z \sim \mathcal{N}(\mu, \sigma^2 I)$.

The reparameterization trick Due to its stochastic nature, this sampling process is inherently non-differentiable. To address this problem, the reparameterization trick is applied: Instead of directly sampling $z \sim \mathcal{N}(\mu, \sigma^2 I)$, we take a sample $\epsilon \sim \mathcal{N}(0, I)$ and use it to calculate $z = \mu + \sigma\epsilon$. By treating ϵ as a constant, differentiation becomes possible again.

The ELBO loss To use the model to generate new data, we want to be able to sample a latent vector z from the prior distribution $p(z)$ and use it to calculate $x' \sim p_\theta(x|z)$. In order to generate high quality data, we need to optimize the parameters θ, such that we maximize the probability that our model generates data from the distribution. That is, we want to maximize $p_\theta(x)$ or, equivalently, maximize $\ln p_\theta(x)$, which is also called the *marginal likelihood* or the *model evidence*. Since $\ln p_\theta(x)$ is intractable, meaning that it cannot be directly optimized, a lower bound of it is optimized instead, called the Evidence Lower BOund (ELBO). In the original paper, this ELBO is derived to be the following:

$$\ln p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}[\ln p_\theta(x|z)] - D_{KL}[q_\theta(z|x) \| p(z)] \tag{2.1}$$

Where $q_\phi(z|x)$ is our encoder's approximation of the posterior with parameters ϕ and $p_\theta(x|z)$, $p_\theta(x|z)$ is the decoder's likelihood function with pa-

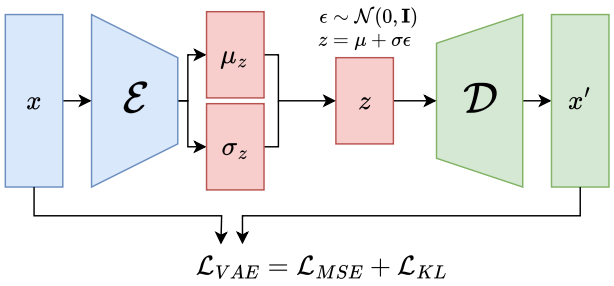


Figure 2.5: **The Variational Auto-Encoder.** The encoder $\mathcal{E}$ learns the conversion from an input x to a mean μ_z and standard deviation σ_z, which together describe a normal distribution $\mathcal{N}(\mu_z, \sigma_z^2)$. Using the reparameterization trick, a latent vector $z \sim \mathcal{N}(\mu_z, \sigma_z^2)$ can be sampled. The decoder model $\mathcal{D}$ in turn learns a conversion from the latent space back to the image space (x'). The loss function now includes the KL divergence loss to ensure z is normally distributed. During inference, $\mathcal{D}$ can be used to generate new samples.

rameters θ and $p(z)$ is the prior distribution. The first component of this equation, representing the likelihood of reconstructing the original data from z given the model, can be estimated with a reconstruction error. The lower the reconstruction error, the higher the likelihood of our model. The second part of the equation is simply the Kullback-Leibler (KL) [37] divergence between the latent space of our model $q(z|x)$ and the desired prior over our latent space $p(z)$. Since the latent space should be normally distributed, this last term is the KL divergence between the latent space of the model and the Gaussian distribution. Based on this, the following loss function for the VAE model can be derived:

$$\mathcal{L}_{VAE} = \mathcal{L}_{ELBO} = \mathcal{L}_{recon} + \mathcal{L}_{KL} \tag{2.2}$$

Because of the likelihood-based training of VAEs, the resulting model has excellent mode coverage, meaning that the output it produces follows the distribution of the original training data, without under- or over-representing certain modes. However, a side effect of this likelihood-based training is that the model predicts the expected value of each pixel, instead of the most likely value. This has the result that, when multiple values are likely, the output is an average over possible values. This can lead to blurry results.

Vector Quantized Variational Auto-Encoders

VQ-VAEs [61] adapt the traditional VAE architecture by adding vector-quantization. During training of this framework, the encoder $\mathcal{E}$ maps an

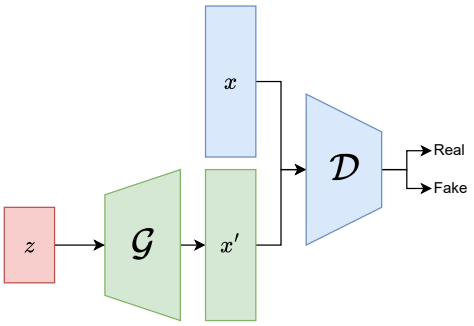


Figure 2.6: **The Generative Adversarial Network.** GANs consist of two models: a generator $\mathcal{G}$ that creates fake samples x' from random noise z, and a discriminator $\mathcal{D}$ that distinguishes between real samples x and fake samples x'. The two models are trained jointly through adversarial optimization, where the generator loss is derived from the discriminator's classification performance on generated samples.

input volume x to a latent representation $z_e(x) \in \mathbb{R}^{H \times W \times D \times C}$ which is then quantized by replacing each feature vector with its nearest neighbor e_k from a codebook $\mathcal{Z}$ of size K. The resulting quantized embedding $z_q(x) \in \mathbb{R}^{H \times W \times D}$ is defined as:

$$z_q(x)_{i,j,k} = \text{argmin}_{e_k} \|e_k - z_e(x)_{i,j,k}\|_2 \tag{2.3}$$

During training, a codebook loss $\mathcal{L}_{codebook} = \|\text{sg}[z_e(x)] - z_q\|_2^2$ and commitment loss $\mathcal{L}_{commit} = \|z_e(x) - \text{sg}[z_q]\|_2^2$ are added, where the sg operator prevents gradient propagation through these terms. These loss functions promote alignment between the codebook vectors and the encoder outputs.

2.2.2 Generative Adversarial Networks

Generative Adversarial Networks [23] (GANs) take a different approach from auto-encoders. Instead of working with a single network consisting of an encoder and a decoder and minimizing the ELBO loss for both components jointly, they consist of two separate models: a generator, and a discriminator. The generator model $\mathcal{G}$ that learns to generate a sample x' from a random noise sample z, while the discriminator $\mathcal{D}$ is trained to distinguish between real data samples x and fake samples x' produced by the generator.

Compared to VAEs, images produces by GANs are very realistic and include sharp details at the expense of mode coverage. However, they are notoriously difficult to train, since they are susceptible to mode collapse where the generator learns to generate one specific type of example that can fool the discriminator. They are also prone to vanishing gradient problems.

Vector-Quantized Generative Adversarial Networks

Vector-Quantized Generative Adversarial Networks (VQ-GANs) [18] combine the strengths of VQ-VAEs and GANs. To improve the realism and quality of the reconstructed images, they incorporate adversarial training by jointly learning a discriminator alongside the autoencoder.

Lastly, alongside the traditional reconstruction loss $\mathcal{L}_{rec} = \|x - \hat{x}\|$ present in regular (VQ)-VAEs, an additional perceptual loss $\mathcal{L}_{perc} = \|\phi(x) - \phi(\hat{x})\|_2^2$ is used, where ϕ represents a feature extractor.

2.2.3 Diffusion Models

Diffusion Models were first introduced by Sohl-Dickstein et al. in 2015 [57] and were heavily inspired by the physical diffusion process where material moves from regions of a high density to regions of a lower density over time. In this process, we start with the material following a very specific, complex distribution (all the material being concentrated in places of high density), and for each timestep t this distribution shifts towards a simpler, more uniform distribution (all the material being equally distributed with uniform density). With Diffusion Models, we perform a similar process: starting with the distribution of our data $p(x_0)$, we apply a transformation $q(x_t|x_{t-1})$ for each timestep t out of T total time steps until we reach the Gaussian distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$. The transformation we apply is the addition of Gaussian noise according to a variance schedule $\beta_1, \ldots, \beta_T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2.4}$$

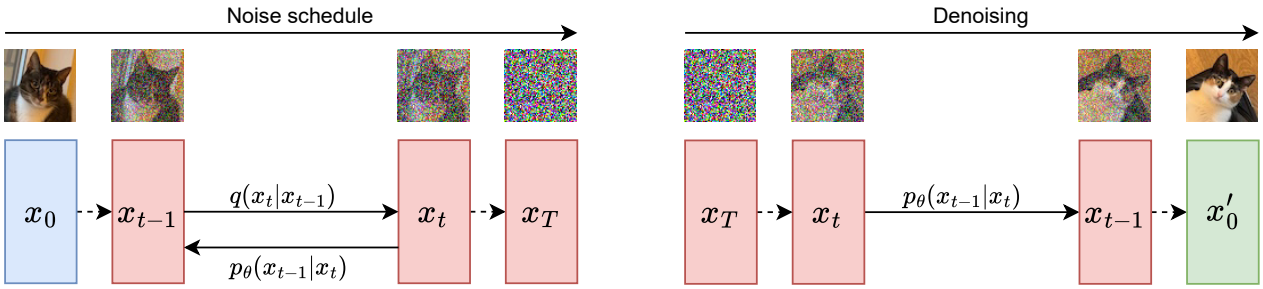
We multiply the mean by $\sqrt{1-\beta_t}$ in order to ensure that x_t has the same variance as x_{t-1}.

During training, we want to optimize our model to provide the highest probability on the training data $p_\theta(x_0)$, which is equivalent to minimizing the expectation of the negative log likelihood: $\mathbb{E}[-\log p_\theta(x_0)]$. Since this problem itself is intractable, we instead minimize the lower bound of the negative log-likelihood, bringing us to an ELBO loss that is formulated as follows:

$$L_{ELBO} = \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] = \mathbb{E}_q \left[-\log p(x_T) - \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \tag{2.5}$$

During training, the model is taught to predict $p_\theta(x_{t-1}|x_t)$. To improve training efficiency, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ can be pre-computed for every timestep t. Where $\alpha_t = 1 - \beta_t$. This allows us to calculate $q(x_t|x_0)$ in one step:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{2.6}$$



(a) Training

(b) Sampling

Figure 2.7: **DDPM training and sampling process.** a) During training, noise is progressively added to real images, and the model p_θ learns to predict this noise. b) During sampling, the process is reversed: starting from pure noise, the model iteratively removes predicted noise to generate new images.

Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs), introduced in [29], made two major improvements to the Diffusion architecture.

First, the diffusion process was reframed as an iterative denoising task. Rather than directly modeling $p_\theta(x_{t-1}|x_t)$, they train a model $\epsilon_\theta(x_t, t)$ that predicts the noise ϵ_t added at timestep t. The simplified loss function then becomes:

$$L_{simple} = \mathbb{E}_{t,x_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2] \tag{2.7}$$

During sampling, we start with pure noise $x_T \sim \mathcal{N}(0, I)$ and iteratively apply our learned denoising model (see Figure 2.7b):

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)\right) + \sigma_t z \tag{2.8}$$

Where $z \sim \mathcal{N}(0, I)$ if $t > 1$ and $z = 0$ if $t = 1$ and $\sigma_t = \sqrt{\beta_t}$ is a variance parameter.

In total, this generation process takes T denoising steps, making it quite computationally expensive. Despite these trade-offs, diffusion models outperform GANs and VAEs in image quality while having a high mode coverage [15].

Denoising Diffusion Implicit Models

While DDPMs outperformed existing methods in terms of image fidelity and mode coverage, they still had significant limitations. Firstly, their inherently Markovian nature means that to create a sample, the model has to

be run for T time steps, making inference very costly. Secondly, due to sampling being a stochastic process, the process is inherently non-deterministic, meaning that if you use the same sample from the latent space, the resulting output will be different each time. This has the additional side effect that the latent space is not semantically meaningful and smooth interpolation between samples is not possible.

To alleviate these limitations, Denoising Diffusion Implicit Models (DDIMs) were introduced [58]. DDIMs redefine the sampling method seen in Equation 2.8 as follows:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t \epsilon_t \tag{2.9}$$

In this equation, different values for σ_t can be used to elicit different sampling behaviors. Using $\sigma_t = \sqrt{\frac{(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}}\sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$, the sampling behavior is exactly the same as regular DDPM sampling. When we set $\sigma_t = 0$, however, sampling becomes entirely deterministic. Note that this method does not require a different training approach, meaning that a model initially trained for DDPM sampling can be used for DDIM sampling without the need for retraining.

2.2.4 Latent Diffusion Models

Latent Diffusion Models (LDMs) [51] were introduced to make diffusion models viable for high-resolution image generation by introducing a two-stage synthesis approach. First, an auto-encoder model is trained to translate between pixel-space and a compressed latent space. Subsequently, a diffusion model is trained to generate samples in the latent space, which can be translated into pixel-space using the decoder model of the auto-encoder. This setup allows the diffusion model to learn semantically meaningful representations, while the auto-encoder ensures visual realism of the final results.

Chapter 3

Related Work

This chapter reviews the current landscape of deep learning in medical imaging, examining state-of-the-art architectures for segmentation and detection tasks across supervised and unsupervised paradigms. We focus particularly on diffusion models for unsupervised medical anomaly detection and position our work within the existing literature.

3.1 Deep Learning in Radiology

Deep learning has proven to be a powerful tool in medical imaging, particularly in Computer-Aided Detection (CAD) or Diagnosis (CADx) systems that assist radiologists in locating and characterizing potential pathologies in medical images. Studies in recent years have shown increasing evidence that deep learning-based systems can achieve similar or improved performance compared to radiologists on detection tasks [39, 27, 53, 5]. It has also been shown in several contexts to outperform traditional risk stratification methods [69, 62] and to be a promising tool for diagnosis and treatment selection [59, 8]. Most importantly, clinical trials have shown that the use of deep learning in radiology can improve patient outcomes while reducing the workload of radiologists [47, 63, 38], solidifying the position of deep learning as a useful and reliable tool in radiology.

3.2 Supervised Methods

Deep learning models in radiology are typically trained using supervised learning, requiring labeled training samples with corresponding annotations. The U-Net [52] has become one of the most influential architectures in medical imaging, utilizing an encoder-decoder structure with skip connections for image-to-image mapping tasks. Originally introduced for medical image segmentation, U-Net has been successfully adapted to diverse applications

including image denoising [41], registration [3], and cross-modal translation [70].

Building upon this foundation, nnU-Net [32] introduced a self-configuring framework that automatically optimizes preprocessing parameters, training procedures, and U-Net architecture for new datasets. This addresses a key challenge in medical imaging: the need for task-specific architectural modifications and hyperparameter tuning. nnU-Net has achieved state-of-the-art performance across numerous medical imaging challenges, including recent superior results on kidney lesion detection [10].

Complementing these medical-specific approaches, general computer vision architectures have also been successfully adapted for medical imaging. The YOLO [49] (You Only Look Once) architecture has been applied to various medical detection tasks [48] and extended to fully volumetric medical images [56], demonstrating the adaptability of established detection frameworks to medical applications.

3.3 Unsupervised Methods

A significant limitation of supervised deep learning methods for medical imaging tasks is the requirement of large annotated training datasets. Detection tasks require annotations such as bounding boxes or point annotations to signify the location and size of medical objects of interest. Segmentation tasks require pixel-level annotations of these objects. Both annotation types are time-consuming to produce and require significant medical expertise.

Unsupervised methods offer a promising alternative by learning from unannotated data. These approaches are particularly valuable in medical imaging where pathological cases are often rare and diverse, making comprehensive annotation challenging.

Variational Autoencoders [36] have emerged as a popular choice for unsupervised anomaly detection in medical imaging due to their ability to learn compact latent representations of normal anatomy [13, 71]. By training on healthy samples, VAEs learn to encode normal anatomical variations in a lower-dimensional latent space. During inference, images that cannot be accurately reconstructed or that produce latent representations far from the learned distribution are flagged as potential anomalies.

An innovative lesion detection approach employs a Maximum-A-Posteriori framework with a Gaussian-Mixture Variational Autoencoder [16] trained solely on healthy data to model the prior distribution of normal anatomy, enabling unsupervised identification of abnormalities [13] on brain MRI. This method leverages the assumption that healthy tissue follows predictable patterns, and deviations from these learned representations indicate potential pathology. The approach demonstrated promising results in detecting various types of lesions without requiring explicit lesion annotations during

training.

Generative Adversarial Networks (GANs) [23] represent another class of unsupervised methods that have shown promise in medical anomaly detection. These approaches typically involve training a generator to produce realistic healthy images while a discriminator learns to distinguish between real and generated samples. The trained generator can then be used to generate healthy reconstructions of potentially pathological images, with differences between the original and reconstructed images highlighting potential anomalies. AnoGAN [54] and its variants have demonstrated this approach's effectiveness across various medical imaging modalities.

3.4 Diffusion Models for Medical Anomaly Detection

More recently, diffusion models have emerged as a promising alternative to VAEs and GANs for medical anomaly detection [46, 67, 68]. The fundamental approach involves training diffusion models on healthy data and using them to reconstruct healthy versions of potentially pathological images during inference. Differences between original and reconstructed images indicate potential anomalies.

Early DDPM-based approaches [46] suffered from inaccurate reconstructions and numerous false positives. Subsequent work demonstrated that DDIM sampling with classifier guidance [67, 20, 2] provides more deterministic and accurate reconstructions, outperforming both naive DDPM approaches and existing reconstruction-based methods. Alternative approaches have explored using simplex noise instead of Gaussian noise to improve reconstruction quality [68, 44], techniques to iteratively focus the reconstruction process only on anomalous regions [6], or using likelihood-based methods to detect anomalies based on the activations of diffusion models [31].

3.5 Research Gap

Despite promising developments, current diffusion-based anomaly detection methods face several significant limitations that restrict their clinical applicability. First, all existing approaches operate on 2D images or individual slices of 3D scans, fundamentally limiting their clinical utility since radiologists routinely analyze 3D volumetric data to understand spatial relationships and contextual information crucial for accurate diagnosis. Processing 3D scans slice-by-slice discards valuable volumetric information and inter-slice dependencies. Limited research exists applying diffusion-based anomaly detection directly to 3D imaging, with only one work comparing Latent Dif-

fusion Models to Latent Transformer Models for out-of-distribution detection [24], a task closely related to anomaly detection.

Second, existing research has predominantly focused on MRI imaging of the brain [67, 68], with limited work done on abdominal imaging [2]. This narrow focus limits generalizability to other anatomical regions that present different challenges in anatomy complexity, contrast characteristics, and pathology presentation.

Finally, current methods have primarily been evaluated against other unsupervised approaches such as VAEs and GANs [67, 68, 6, 46]. While these comparisons demonstrate relative merits among unsupervised techniques, they fail to establish whether diffusion-based methods can compete with state-of-the-art supervised methods that represent the current clinical standard.

Chapter 4

Methods and Materials

In this chapter, we present the methods and materials used in our research. We first introduce the proposed method for kidney anomaly detection using diffusion models, giving an overview of the overall pipeline, the post-processing we use, and the mathematical foundations behind its components. We then describe the datasets used for training and evaluation, including our annotated datasets, as well as our unannotated dataset and the procedure we use to generate pseudo-labels. Lastly, we cover the implementation and training details for our proposed method.

4.1 Overview of the Proposed Method

We propose a novel method based on existing research on diffusion-based anomaly detection and latent diffusion for medical imaging. We implement a pipeline for kidney anomaly detection in 3D CT volumes. In this section, we give an overview of the proposed pipeline, including its components and the underlying mathematical principles.

4.1.1 Anomaly Detection Pipeline

A visual overview of our proposed pipeline is shown in Figure 4.1. Each input CT scan is resampled to 1 mm isotropic resolution. Kidney segmentation masks are obtained using TotalSegmentator [65] (fast preset), and $96 \times 96 \times 128$ mm patches are extracted around each kidney. These patches are encoded into a latent representation using the VQ-GAN encoder (Subsection 2.2.4). Within the latent space, the patch undergoes L steps of the forward (noising) diffusion process followed by L denoising steps using classifier-guided DDPM or DDIM sampling (Subsection 4.1.2). The denoised latent patch is decoded back into pixel space using the VQ-GAN decoder, producing a healthy reconstruction of the original input. Anomaly maps are computed as voxel-wise differences between the original patch and the

reconstruction, resampled to the native resolution, and reconstructed into a full-scan anomaly map.

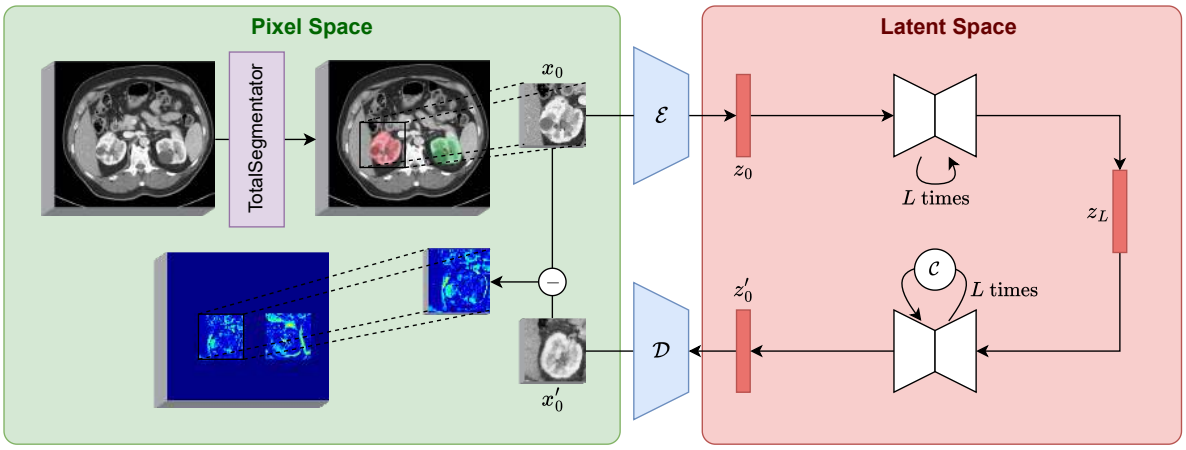


Figure 4.1: **Overview of the proposed anomaly detection pipeline.** Kidney patches, extracted using TotalSegmentator masks, are encoded into a latent space via the encoder of a VQ-GAN. The latent patch undergoes L forward (noising) diffusion steps, followed by L reverse (denoising) steps with classifier guidance. The denoised latent patch is decoded to produce a healthy reconstruction. Subtracting this from the original yields the anomaly map.

Post-Processing

After obtaining the anomaly map, we apply several post-processing steps to convert it into a binary segmentation map and reduce false positives.

First, thresholding is applied to the anomaly map produced by the pipeline to convert it into a binary segmentation map. Similar to [67], the threshold value is calculated by globally applying the Otsu method [45] over all anomaly maps produced for the test set. This method automatically determines the optimal threshold for converting a grayscale image to binary by finding the threshold value that minimizes the weighted sum of within-class variances for the foreground and background pixel groups.

To reduce noise and false positives in the resulting binary map, we first mask the binary segmentation to include only voxels within the kidney regions, as determined by the TotalSegmentator kidney segmentation masks. Morphological opening and closing operations are then applied, followed by a hole filling algorithm. This reduces noise and artifacts generated around the border of the kidney. Connected component analysis is performed and a two-step size filtering approach is applied to remove small artifacts. First, we remove any connected components smaller than 20 voxels in total, which is computationally efficient to calculate. Subsequently, we apply a more

precise geometric constraint by removing lesions with a largest diameter smaller than 3mm in the axial plane, measured by taking the largest side of the smallest fully covering bounding box. This 3mm threshold aligns with the size criteria used for our baseline methods (Subsection 5.3.1), ensuring consistent evaluation across all approaches.

4.1.2 Anomaly Detection Using Diffusion Models

To apply DDPMs to anomaly detection, a noised version x_L of the input is created using the forward process (Equation 2.4), followed by the sampling process described in Equation 2.8 for $T - L$ steps to construct x'_0, which is expected to approximate a healthy version of the original image. Additionally, classifier guidance [15] can be used to explicitly guide the generative process towards healthy reconstructions by replacing ϵ_θ with $\hat{\epsilon}$ defined as:

$$\hat{\epsilon}(x_t) = \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log C_\phi(y|x_t, t) \tag{4.1}$$

where C_ϕ is a classifier on noisy images x_t.

Due to the stochastic nature of the noising process, standard DDPMs have been shown to yield suboptimal results for anomaly detection [67, 68]. To address this limitation, the proposed method compares Denoising Diffusion Implicit Models (DDIMs) [58], which provide a deterministic and reversible sampling process, with DDPMs. In DDIM, the reverse step is formulated as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t) \tag{4.2}$$

As shown in [67], this process can be reversed to calculate x_{t+1} given x_t:

$$x_{t+1} = x_t + \sqrt{\bar{\alpha}_{t+1}}\left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}}\right)x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1}\right)\epsilon_\theta(x_t, t)\right] \tag{4.3}$$

allowing deterministic noising and denoising of an original image.

4.1.3 3D Latent Diffusion

To enable efficient processing of 3D medical images, the proposed method applies latent diffusion [51], where the diffusion process operates in a compressed latent space rather than directly in pixel space, as explained in Subsection 2.2.4.

For this purpose, we adopt the architecture introduced in [35], which was designed for 3D medical image generation. This architecture uses a VQ-GAN Section 2.2.2 as the autoencoder model. During training, adversarial loss is calculated using a slice-level discriminator, as well as a volume-level discriminator. The first five blocks of a pre-trained VGG16 function as the feature extractor used to calculate the perceptual loss.

4.2 Datasets

4.2.1 Supervised Training Data

This subsection describes the annotated datasets obtained from Radboudumc and the KiTS challenges. An overview of the characteristics for each individual dataset, as well as for the combined datasets used to train nnU-Net and nnDetection, is provided in Table 4.1.

Table 4.1: **Characteristics of our supervised training datasets.** This table compares properties of the KiTS (USA) and Radboudumc (Netherlands) datasets, both individually and when combined for training nnU-Net (full dataset, 693 scans) and nnDetection (80% subset, 554 scans). For each dataset, we report the country of origin and number of cases. For per-case metrics (lesion count, mean lesion volume, mean kidney volume including lesions, in-plane resolution, and slice thickness), we report the mean and standard deviation across all cases. In the calculation of lesion and kidney properties, we first remove disconnected artifacts smaller than 2.0mm^3 and 100.0mm^3 for lesions and kidneys, respectively. Lastly, we ignored cases without lesions in the calculations for lesion count and mean volume.

	KiTS	Radboudumc	Radboudumc + KiTS	
			nnU-Net (100%)	nnDetection train (80%)
Country	USA	Netherlands	Mixed	Mixed
Number of scans	479	214	693	554
Number of lesions	2.8 (±2.9)	4.6 (±5.6)	3.2 (±3.8)	3.2 (±3.6)
Lesion volume (10^3 mm^3)	85.5 (±181.6)	17.72 (±84.5)	70.4 (±167.4)	69.4 (±170.3)
Kidney volume (10^3 mm^3)	277.5 (±154.1)	187.0 (±105.3)	249.7 (±146.9)	246.8 (±139.0)
In-plane resolution (mm)	0.80 (±0.11)	0.75 (±0.07)	0.78 (±0.10)	0.78 (±0.10)
Slice thickness (mm)	3.35 (±1.70)	1.22 (±0.37)	2.69 (±1.73)	2.67 (±1.73)

KiTS

The KiTS dataset, made publicly available for the KiTS19 [26], KiTS21 [25], and KiTS23 challenges, contains images and corresponding annotations for 489 CT images belonging to 489 patients. All patients in the dataset underwent a medical procedure for suspected kidney cancer between 2010 and 2022 at USA-based hospitals. The included CT scans are contrast-enhanced scans taken before the procedure. The annotations contain segmentations for the kidney (including non-fat tissue in the hilum), tumors, and cysts. Annotated tumors are masses that were confirmed to be malignant through histopathology. Annotated cysts were confirmed radiologically or histopathologically. Annotations were performed by trainees (including residents, medical students, and undergraduates planning to study medicine), trained by experts

(consisting of radiologists, urologic oncologists, and urologic oncology fellows). In cases of low agreement between trainees, segmentations were further reviewed by the experts.

For our research, we used 479 scans from this dataset. Since we are not interested in subtyping the lesions but only in detecting all lesions, the classes for suspected tumors and cysts were merged into a single lesion class.

Radboudumc

The second supervised training set is a publicly available dataset from Radboudumc [40]. The dataset contains 215 contrast-enhanced CT scans from 215 patients taken at Radboudumc during clinical routines in 2015. 102 of these cases contain kidney abnormalities, including cysts, lesions, masses, metastases, or tumors. The other 113 cases do not contain any kidney abnormalities. The full patient cohort (which includes the dataset described in Subsection 4.2.3 as well as the one described in this paragraph) has an average age of 60, ranging from 22 to 84 years old, and consists of 56% males. The cohort also contains cases where a nephrectomy (kidney removal) has already been performed. This dataset contains 17 cases where a left nephrectomy was performed and 18 cases with a right nephrectomy.

Annotations were performed by four medical students, trained and, in ambiguous cases, assisted by an experienced radiologist. Segmentations include the full kidney and the abnormalities. All abnormalities were segmented as a single class. The urine collection system, kidney hilum, and lesions within them were excluded from both segmentations.

In this study, we use 214 scans from this dataset.

Combined datasets

We follow the same procedure for combining the datasets as was used for the pre-trained nnU-Net [10] that we use as a supervised baseline. This includes merging the labels for tumors and cysts into a single abnormality label and accepting the inconsistency in kidney hilum inclusion between both datasets. We use 80% of this dataset to train our baseline nnDetection model and to perform hyperparameter search for our proposed methods. We also reserve 20% as a holdout set for our proposed methods and the nnDetection model, balanced to ensure an equal proportion of samples from KiTS and Radboudumc in both splits.

The dataset properties of the full combined dataset, as well as the 80% training split can be seen in Table 4.1.

4.2.2 Supervised Test Sets

We use two test sets to evaluate our proposed methods, as well as the supervised methods. Table 4.2 describes the properties of these test sets.

nnDetection test set

The first test set is the 20% of cases that were held out during training of our nnDetection model (see Section 4.2.1), consisting of 139 cases. This dataset follows a similar distribution of properties as the nnDetection training set. Since these cases were included in the training dataset of the nnU-Net baseline, we excluded this baseline model from evaluation on this test set.

Radboudumc private test set

We use the same private test set as in [10]. This set consists of 50 cases from Radboudumc [40] and follows the same acquisition and annotation procedures and has the same distribution of properties as the dataset described in Section 4.2.1. In this dataset, 2 and 6 patients have undergone left and right nephrectomy, respectively. The test set contains 30 cases with kidney abnormalities, which were used to evaluate segmentation and detection performance of our proposed methods, as well as our supervised baselines.

Table 4.2: **Characteristics of our supervised test datasets.** The first test set is the hold-out set of the nnDetection training data and is a combination of scans from KiTS (Netherlands) and Radboudumc (NL), described in Table 4.1. The second test set is a private test set from Radboudumc, where the full set (used to evaluate the classifier) and the subset of unhealthy cases (used to evaluate segmentation and detection) are shown. For each dataset, we report the country of origin and number of cases. For per-case metrics (lesion count, mean lesion volume, mean kidney volume including lesions, in-plane resolution, and slice thickness), we report the mean and standard deviation across all cases. In the calculation of lesion and kidney properties, we first remove disconnected artifacts smaller than 2.0mm^3 and 100.0mm^3 for lesions and kidneys, respectively. Lastly, we ignored cases without lesions in the calculations for lesion count and mean volume.

	Radboudumc + KiTS	Radboudumc private test set	
	nnDetection test (20%)	All	Unhealthy
Country	Mixed	Netherlands	Netherlands
Number of scans	139	50	30
Number of lesions	3.2 (±4.5)	3.8 (±2.7)	3.8 (±2.7)
Lesion volume (10^3 mm^3)	74.6 (±155.2)	15.5 (±46.9)	15.5 (±46.9)
Kidney volume (10^3 mm^3)	260.3 (±174.3)	182.8 (±63.9)	191.0 (±74.5)
In-plane resolution (mm)	0.78 (±0.10)	0.76 (±0.06)	0.75 (±0.07)
Slice thickness (mm)	2.77 (±1.75)	1.17 (±0.38)	1.22 (±0.37)

4.2.3 Unsupervised Training Data and Pseudo-Labels

The VQ-GAN, DDIM, DDPM, and classifier models are trained on a private dataset from Radboudumc. Scans from clinical routine between 2008 and 2021 were selected for which the institutional review board waived the need for informed consent due to the study's retrospective design and the pseudonymization of data. From this set, we selected all contrast-enhanced thorax-abdomen/abdomen CT scans with a slice thickness ≤ 1 mm. This resulted in $8,377$ scans from $7,571$ studies from $6,800$ patients, with $5,095$ left and $5,099$ right kidneys extracted. Pseudo-labels were derived from radiology reports: kidneys reported as having lesions, cysts, or tumors were labeled unhealthy; those described as normal or unmentioned were labeled healthy. Kidneys with stones, calcifications, necrosis, atrophy, or prior removal were excluded. If the radiology report was found to be empty or contained less than 10 sentences, the kidney was labeled as unknown. The data selection and pseudo-label generation processes are summarized in Figure 4.2.

We used all non-excluded kidneys to train the VQ-GAN. Our diffusion models were trained only on kidneys labeled healthy, while both healthy and unhealthy kidneys were included in the training data for our classifier.

4.3 Implementation and Training Details

The base architecture for our latent diffusion method is based on the VQ-GAN and denoising U-Net models from [35]. We implemented DDPM and DDIM samplers from scratch that support classifier guidance during sampling using our classifier model. This section details the implementation and training details for each component of our overall pipeline. For a comprehensive overview of all hyperparameters used, see Appendix A.

4.3.1 VQ-GAN

We adopted the VQ-GAN architecture from [35] without adaptations. The model was trained on all non-excluded kidney patches described in Subsection 4.2.3 for $100,000$ iterations with a batch size of 4. Training required less than 1 day on an A100 GPU.

4.3.2 DDPM/DDIM

We utilized the denoising U-Net architecture from [35] and trained it for $250,000$ iterations on the healthy kidney patches described in Subsection 4.2.3. We modified the number of timesteps from 300 to 1000 and implemented early stopping to select the model with the lowest validation loss. The model was trained for $250,000$ iterations, with batch size 40, on an A100 GPU, requiring 3 days.

4.3.3 Classifier

For the classifier, we used the downsampling path of the denoising U-Net described in the previous subsection and added a classification head consisting of 2 linear layers. The model was trained on a balanced dataset comprising equal numbers of healthy and unhealthy kidneys ($1,162$ each) using pseudo-labels (Subsection 4.2.3), with an 80/20 train/validation split. Due to extreme overfitting problems encountered during training, we applied heavy regularization techniques consisting of weight decay, dropout layers, and data augmentation. We also froze the first layers of the U-Net backbone. The regularization techniques, along with other hyperparameters, are detailed in Appendix A. Training for 100 epochs with a batch size of 32 required < 2 hours on an RTX2080 Ti. Early stopping selected the model with the highest validation AUC of 0.72.

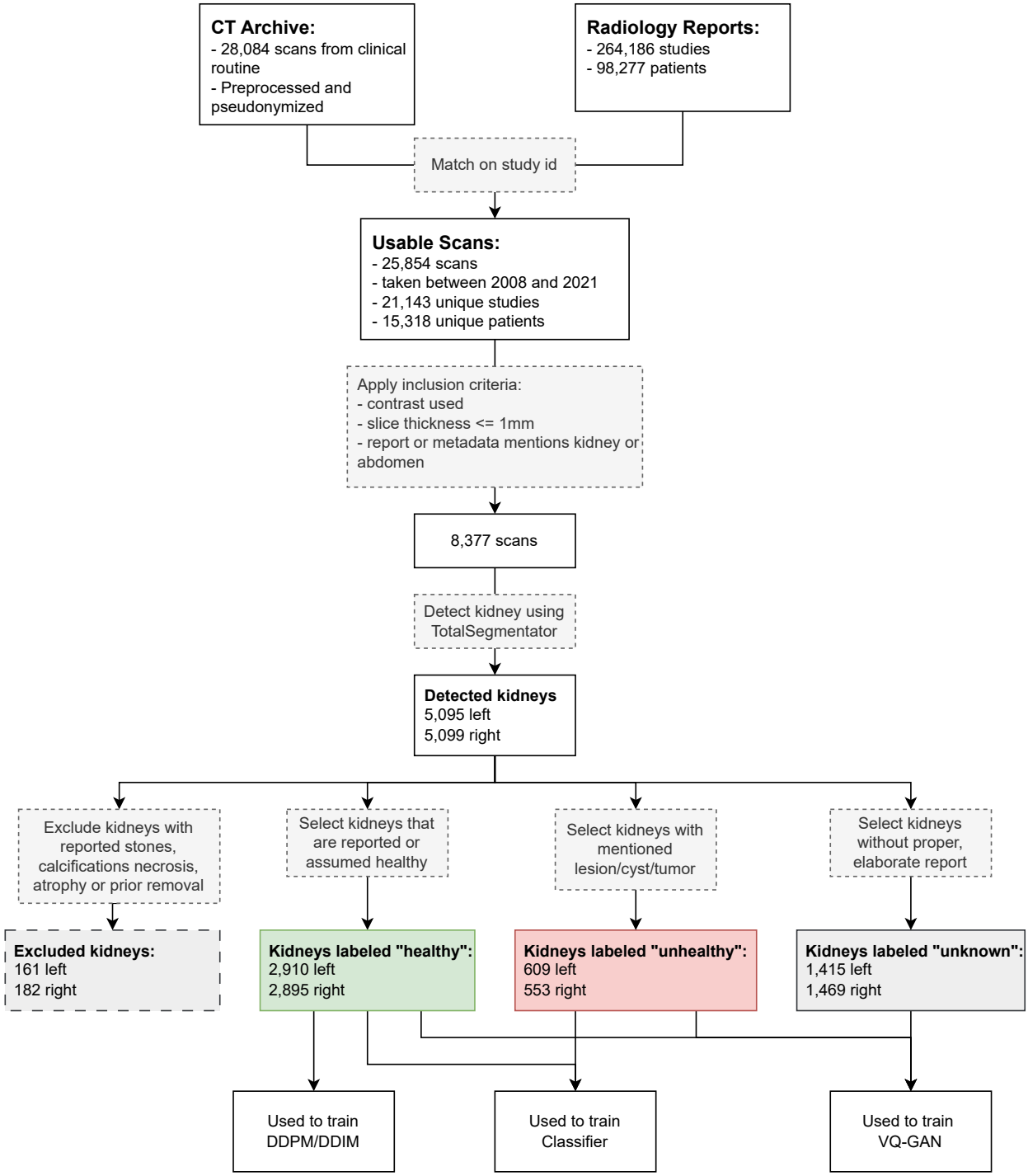


Figure 4.2: **Flowchart of the data selection process and generation of pseudo-labels.** We start by selecting usable scans that have an associated radiology report. We then select scans based on our inclusion criteria and extract the kidneys using TotalSegmentator. Based on the associated radiology report, we exclude the kidney or label them as *healthy*, *unhealthy*, or *unknown*. The DDPM/DDIM is trained only on kidneys labeled *healthy*, the classifier on kidneys labeled *healthy* or *unhealthy*, and the VQ-GAN is trained on all kidneys.

Chapter 5

Experiments

This chapter details the experimental setup used to answer the research questions. First, a glossary of all evaluation metrics and their definitions is provided. We then describe the hyperparameter optimization process used to identify optimal model configurations for subsequent experiments. Following this, the evaluation methodology for assessing overall performance of the proposed method and its comparison to supervised benchmarks is explained, including evaluation across different lesion sizes and qualitative analysis. Finally, the experiments conducted to assess the performance of individual pipeline components are presented.

5.1 Evaluation Metrics

Throughout this chapter, several standard metrics are used to evaluate segmentation, detection, generation, and classification performance. These metrics are defined here to ensure clarity in the interpretation of results.

5.1.1 Segmentation Metrics

For segmentation evaluation, the **Dice Similarity Coefficient (DSC)**, also known as the Dice score, is used. The DSC measures the overlap between predicted and ground truth segmentation masks:

$$\text{DSC} = \frac{2 \times |P \cap G|}{|P| + |G|} \tag{5.1}$$

Where $|P \cap G|$ represents the number of voxels in the intersection of the predicted segmentation P and the ground truth segmentation G, $|P|$ represents the number of voxels in the predicted segmentation, and $|G|$ represents the number of voxels in the ground truth segmentation. The DSC ranges from 0 (no overlap) to 1 (perfect overlap), with higher values indicating better segmentation performance.

5.1.2 Detection Metrics

For detection evaluation, three complementary metrics that provide different perspectives on model performance are used:

Precision measures the fraction of predicted lesions that correspond to actual lesions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.2}$$

Recall, also known as sensitivity, measures the fraction of actual lesions that are successfully detected:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.3}$$

F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

For detection evaluation, a predicted lesion is considered a true positive if it has an Intersection over Union (IoU) above a certain threshold with a ground truth lesion. IoU is defined as the area of overlap between the predicted and ground truth lesions divided by the area of their union. When multiple predictions overlap with the same ground truth lesion, only the prediction with the highest IoU is considered a true positive.

Free-Response Receiver Operating Characteristic (FROC) curves are used to visualize detection performance by plotting sensitivity (recall) against the average number of false positives per scan across different confidence thresholds. FROC curves are particularly useful for medical detection tasks as they directly show the trade-off between sensitivity and false positive rate, which is clinically relevant.

5.1.3 Classification Metrics

For evaluating classifier performance, the **Area Under the Receiver Operating Characteristic Curve (AUC or AUROC)** is used. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) across all possible classification thresholds. The AUC summarizes the classifier's ability to distinguish between classes, with values typically ranging from 0.5 (random chance) to 1.0 (perfect classification).

5.1.4 Generative Model Metrics

For evaluating the diversity of generated samples from diffusion models, the **Multi-Scale Structural Similarity Index Measure (MS-SSIM)** [64]

is used. MS-SSIM computes structural similarity between two images at multiple scales by iteratively low-pass filtering and downsampling the image to create multiple scales, then combining luminance, contrast, and structure comparisons at each scale. The metric ranges from 0 to 1, where 1 represents perfect similarity. For diversity evaluation, the mean pairwise MS-SSIM across all generated images is reported. Lower MS-SSIM values indicate greater structural diversity between samples, while higher values suggest more similarity and less diversity.

5.2 Hyperparameter Optimization

Hyperparameter optimization is performed to find the optimal noise level (L) and classifier guidance strength (s). Since the two sampling methods (DDPM, DDIM) are not expected to share the same optimal hyperparameter configuration, optimization is performed for each method individually. Optimal hyperparameters are determined through a grid search over the values specified in Table 5.1a.

The dataset used for optimization is identical to that used for training the nnDetection baseline, consisting of data from KiTS and Radboudumc, as described in Subsection 4.2.1. The optimization metric is the average DSC score over all cases. For computational efficiency, no post-processing is applied during optimization. The threshold for DSC score evaluation is calculated by applying the Otsu method [45] globally over all predicted segmentation maps.

Table 5.1b shows the optimal set of hyperparameters found for each method. These values are used for the rest of the experiments in this chapter. Figure 5.1 shows the DSC score reached by each method at different hyperparameter configurations.

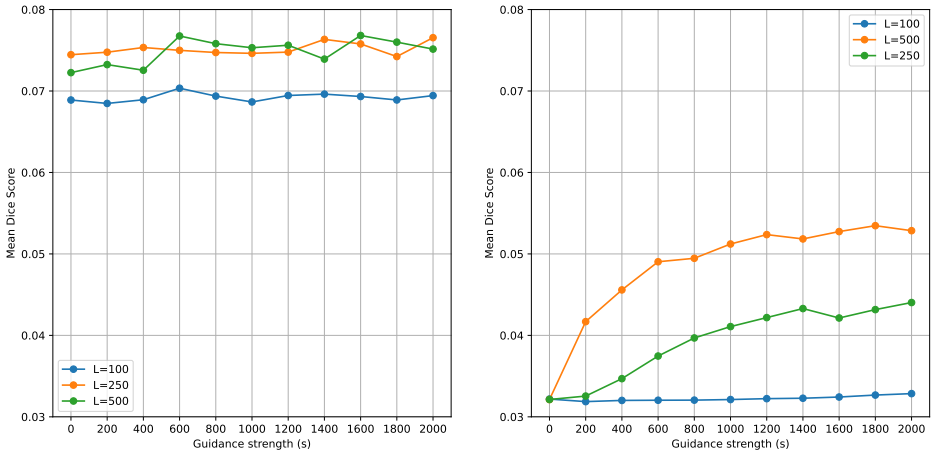
Table 5.1: **Hyperparameter optimization** (a) shows the evaluated values for both sampling methods. (b) shows the hyperparameter configuration that reached the highest DSC score for each sampling method.

(a) Values evaluated.

Parameter	Values
noise level (L)	$[100, 250, 500]$
guidance strength (s)	$[0, 200, \cdots, 2000]$

(b) Optimal values found.

	DDPM	DDIM
L	500	500
s	1600	1800



(a) DDPM (b) DDIM

Figure 5.1: **Evaluation of segmentation performance at different hyperparameter configurations for both models.** The figure shows the segmentation performance (measured in DSC) of the DDPM-based method (a) and the DDIM-based method (b). Each subplot shows the Dice score plotted against the guidance strength for three different noise levels. Within a plot, the point with the highest Dice value shows the best hyperparameter configuration for that model.

5.3 Main Pipeline Evaluation

After finding the optimal hyperparameters for each model, as described in the previous section, we evaluate the performance of the full pipeline. A quantitative evaluation of the segmentation and detection performance of the proposed methods is performed first, comparing them to each other and to the supervised baselines. Subsequently, a qualitative evaluation of the reconstruction and detection performance is conducted by examining representative cases.

5.3.1 Supervised Baselines

We compare our weakly supervised approach to two state-of-the-art supervised baselines. One model is used as-is from prior work and serves as a segmentation and detection benchmark, while the other is trained from scratch to serve as a detection-specific benchmark.

nnU-Net The first baseline is a pre-trained nnU-Net [32] for kidney and kidney lesion segmentation [10]. The method was trained using 5-fold cross-validation on the dataset described in Subsection 4.2.1. The model produces a semantic segmentation map with classes for the background, kidney, and

lesion. The original authors apply post-processing to remove lesions below 3mm in diameter and lesions that are not connected to the kidney and are smaller than 100cm^3. Connected component analysis is then applied to produce an instance segmentation map to evaluate detection performance.

nnDetection For detection benchmarking, nnDetection [4] was trained on kidney lesion detection, using 5-fold cross-validation. The bounding box predictions produced by the model are used, as instance segmentation maps are not officially supported by the framework. 80% of the dataset described in Subsection 4.2.1 was used for training, allowing the remaining 20% to serve as a hold-out test set. The split was balanced on dataset origin, ensuring the same proportion of Radboudumc and KiTS scans in the training data and test data.

We used the same training configuration as used by the original authors for benchmarking on the KiTS19 challenge [26], including only lesions $\geq$ 3mm in diameter in the training data.

Without post-processing, our trained nnDetection model produced many false positives. These primarily presented as multiple overlapping predictions for the same lesion, especially for larger lesions that spanned multiple anchor points of the architecture, or lesions with elongated or curved morphologies. Additionally, some false positives appeared as detections outside the kidney boundaries. We evaluated several strategies to eliminate overlapping predictions, including: (1) merging predictions with significant overlap by removing those with lower confidence scores, and (2) computing weighted averages of overlapping bounding box coordinates. The confidence-based removal strategy was found to be more effective at removing false positives, and we use this method in our post-processing. As a final post-processing step, we applied anatomical constraints by removing any detections that fell outside the kidney boundaries, as determined by the segmentation masks from the nnU-Net baseline (Section 5.3.1).

5.3.2 Segmentation and Detection Evaluation

We perform a quantitative analysis of our proposed methods on the test sets described in Subsection 4.2.2. The threshold to convert the anomaly map produced by the proposed methods into a binary segmentation mask is calculated by globally applying the Otsu method [45] over all predictions for each dataset. For nnU-Net, the class with the highest probability for each voxel is selected, and for nnDetection, only bounding box probabilities with a confidence score > 0.5 are included.

To evaluate segmentation performance and determine whether DDIM sampling outperforms DDPM sampling (**RQ1a**) and to compare our proposed method against the nnU-Net baseline (**RQ2b**), the segmentation performance of the proposed methods and the nnU-Net baseline is evaluated.

The mean DSC score, along with the standard deviation, are reported.

For detection evaluation, we assess whether DDIM sampling outperforms DDPM sampling (**RQ1b**) and compare our proposed methods against the nnDetection baseline (**RQ2c**) on both datasets. We also compare the detection performance of our nnDetection baseline against the nnU-Net baseline on the Radboudumc test set, answering **RQ2a**. Predicted lesions are computed by performing connected component analysis over the binary segmentation map. True positives are counted if a lesion has IoU ≥ 0.2 with a ground truth lesion. When multiple predictions overlap with the same ground truth, only the one with the highest IoU is counted. A relatively low IoU threshold of 0.2 was selected to account for the inherent difficulty in precisely matching 3D lesion segmentations with irregular shapes, compared to the simpler geometric matching of bounding boxes. The mean precision, recall, and F1 scores, along with their standard deviations, are reported.

For detection specifically, a FROC curve is plotted for nnDetection, since it can easily be evaluated at a continuous range of thresholds. The proposed methods and nnU-Net, however, are evaluated at a single operating point each due to the computational infeasibility of evaluating the complete post-processing pipeline across multiple thresholds. Therefore, we chose specific operating points: for the proposed methods, the Otsu method [45] is applied globally over all predictions per dataset and the resulting threshold is applied. For nnU-Net, the class with the highest probability for each voxel is selected.

All detection results include 95% confidence intervals, calculated using bootstrapping with 1000 bootstraps.

5.3.3 Evaluation Across Lesion Sizes

To examine how our proposed methods perform across different lesion size ranges (**RQ1c**), we evaluate the performance of our proposed methods across different lesion size ranges. Lesion sizes are categorized according to the T (tumor size) stage distinctions of the TNM staging system (Section 2.1.2). Table 5.2 shows the distribution of lesions per stage for each test set.

The Radboudumc test set shows a highly skewed distribution, with the vast majority of lesions (93 out of 98, or 95%) falling within stage T1a. This severe imbalance makes meaningful size-based evaluation unfeasible on this test set. Therefore, size evaluation is restricted to the nnDetection test set, which has a larger sample size and exhibits a slightly more balanced distribution across size categories.

For the nnDetection test set, the majority of lesions (285 out of 355, or 80%) still fall within the T1a size range (< 4cm). To enable more granular analysis within this dominant category, we subdivide T1a into two ranges: T1ai (< 2cm) and T1aii ($2-4$cm). Additionally, due to the small sample sizes of stages T2a ($n = 15$) and T2b ($n = 16$), we combine these into a

Table 5.2: **Distribution of lesions by size range across test sets.** Size ranges are based on the T stage distinctions of the TNM staging system, with custom subdivisions T1ai and T1aii introduced for the standard T1a stage. The size ranges and test set used for size-based performance evaluation are highlighted in bold.

T Stage	Size Range (cm)	**nnDetection Test**	Radboudumc Test
All	—	355	98
T1	< 7	342	94
T1a	< 4	285	93
T1ai	< 2	213	80
T1aii	2–4	72	13
T1b	4–7	39	1
T2	> 7	31	4
T2a	7–10	15	3
T2b	> 10	16	1

single T2 category (> 7cm). The final size categories used for evaluation are highlighted in bold in Table 5.2.

For evaluation within each size range, the following methodology is applied: First, ground truth lesions are filtered to include only those whose diameter falls within the target size range. Lesion diameter is measured as the largest dimension of the smallest fully enclosing bounding box in the axial plane. Second, any predictions that both (a) do not match a ground truth lesion (IoU ≤ 0.2) and (b) fall outside the target size range when measured using the same diameter calculation method are excluded. Finally, performance metrics for the remaining filtered ground truths and predictions are calculated using the identical methodology described in Subsection 5.3.2. Cases that have no ground truth lesions in the evaluated size range are excluded from evaluation.

5.3.4 Qualitative Analysis

Besides the quantitative analysis of the pipeline described in the previous subsections, we also perform a qualitative analysis. For this evaluation, we select informative examples and visualize the reconstructions and anomaly maps generated by the methods. We investigate examples with different lesion sizes, lesion locations, successful segmentations, and unsuccessful segmentations.

5.4 Sub-Component Evaluation

To get insights on the contribution of the sub-components to the performance of the overall pipeline, we evaluate each component individually. First, the performance of the latent diffusion model itself is examined by evaluating its image generation quality and diversity. The performance of the classifier is also evaluated. Lastly, the accuracy of TotalSegmentator in finding the regions of interest for the diffusion model is assessed.

5.4.1 Latent Diffusion Model Evaluation

We aim to answer **RQ3a** by evaluating our latent diffusion model. Typically, generative models are evaluated on the diversity and fidelity of generated samples. To quantify diversity, we use MS-SSIM. The metric is calculated over the full training dataset and over $1,000$ generated samples. Comparing the MS-SSIM distributions indicates the degree of diversity achieved by our model with respect to the training data.

For qualitative evaluation of the latent diffusion model, we analyze a subset of representative examples generated using both DDPM and DDIM sampling. We also select a subset of real samples for comparison.

5.4.2 Classifier Evaluation

To answer **RQ3b**, we evaluate the performance of the classifier on a combined test set consisting of kidneys from the Radboudumc training set (Section 4.2.1) and the private Radboudumc test set (Section 4.2.2). In contrast to the KiTS dataset, both of the Radboudumc datasets originate from a similar distribution and contain both healthy and unhealthy cases. For each case in the test set, both kidneys are segmented using the ground truth segmentation mask, a patch around the kidney is extracted and labeled healthy if no lesions are present in the ground truth label of the patch, and labeled unhealthy if lesions are present. This results in 486 kidneys extracted from 264 different cases, with an equal split of healthy and unhealthy cases. A more detailed overview can be seen in Table 5.3.

To evaluate the classifier performance, each kidney patch is first encoded into the latent space using the trained VQ-GAN. Three versions of the encoded patch with noise levels 0, 250, and 500 are then created, and the classifier is run for each of these versions. Evaluation is performed by generating ROC curves for the results for each noise level. The AUC for these curves is then calculated.

Additionally, the averages for these three different noise levels are calculated, and a single ROC curve for these averages is created. To quantify the expected range of these values, 95% confidence intervals are calculated and visualized. These intervals are calculated using bootstrapping with 1000

bootstraps.

The results of the experiments described in this subsection can be seen in Subsection 6.2.2.

Table 5.3: **Overview of kidney patches used to evaluate the performance of our classifier.** The dataset consists of our Radboudumc training set (Section 4.2.1) and our private Radboudumc test set (Section 4.2.2). We show the number of left and right kidneys present in the combined dataset, as well as the number of healthy and unhealthy kidneys, based on ground truth annotations.

	Healthy	Unhealthy	Total
Left Kidneys	126	119	245
Right Kidneys	117	124	241
Total	243	243	486

5.4.3 TotalSegmentator Evaluation

In order to answer **RQ3c**, we investigate the performance of TotalSegmentator on both test sets, reporting the mean precision, recall, F1 score, and accuracy over all cases, along with 95% confidence intervals, calculated with 1000 bootstraps. We maintain the same detection criteria as in Subsection 5.3.2, counting only predictions that have an IoU ≥ 0.2 with a ground truth.

5.5 Statistical Analysis Plan

We test the superiority of our DDIM-based method over our DDPM-based method (**RQ1b**) on both test sets described in Subsection 4.2.2. Additionally, we test the superiority (Subsection 5.5.1) of our nnDetection baseline over the nnU-Net baseline (**RQ2a**) on the Radboudumc test set. Subsequently, we test the non-inferiority (Subsection 5.5.2) of both our proposed methods (DDIM and DDPM) over nnDetection (**RQ2c**) on both test sets. For each comparison, the metric we test is the F1 score.

In total, 3 superiority tests and 4 non-inferiority tests are performed for a total of 7 tests. To correct for multiple-testing, Holm-Bonferroni correction [30] is applied, with a total $\alpha = 0.05$. This method ranks all p-values from smallest to largest and compares each one to an increasingly lenient threshold α/n, where $n = 7$ for our test with the highest p-value and $n = 1$ for our test with the lowest p-value. For each test, we report the p-value, corrected α and the effect size in terms of median difference in F1 score.

All tests employ a paired design where each method is evaluated on identical test sets. All statistical testing is done using Python version `3.13.1` and scipy version `1.16.0`.

5.5.1 Superiority Testing

The superiority of method A over B is determined by comparing the median differences against zero using the following null-hypothesis (H_0) and alternative hypothesis (H_1):

- $H_0 : \text{median}(F1_A - F1_B) \leq 0$.
- $H_1 : \text{median}(F1_A - F1_B) > 0$.

A p-value is calculated using the Wilcoxon signed-rank test [66] with a one-tailed alternative hypothesis. This is a non-parametric test that does not assume normal distribution of the test data. The alternative hypothesis (superiority of A over B) is accepted if $p < \alpha$.

5.5.2 Non-Inferiority Testing

Non-inferiority is determined using the same test we use for superiority, but with different null and alternative hypotheses:

- $H_0 : \text{median}(F1_A - F1_B) \leq -\delta$.
- $H_1 : \text{median}(F1_A - F1_B) > -\delta$.

Where $\delta = 0.05$ is the non-inferiority margin, such that A is only considered inferior to B if the comparative performance is worse than $-\delta$. The test is implemented by transforming the differences as $(F1_A - F1_B + \delta)$ and applying the Wilcoxon signed-rank test to determine if the median of the transformed differences is significantly greater than zero.

Chapter 6

Results

This chapter presents the results obtained from the experiments described in Chapter 5. The chapter is structured as follows: first, the performance evaluation of the proposed pipeline is presented and compared against supervised baseline methods. This section includes a comprehensive size-stratified analysis and qualitative assessment of model predictions. Subsequently, the evaluation results for the individual pipeline components are reported and analyzed.

6.1 Main Pipeline

6.1.1 Detection and Segmentation Performance

We evaluate the segmentation and detection performance of our proposed methods as described in Subsection 5.3.2. The results are reported in Table 6.1. From our results, it is clear that our DDPM-based method outperforms our DDIM-based method on most metrics. For segmentation, our DDPM-based method reached a DSC of 0.12 compared to 0.07 reached by our DDIM-based method on our nnDetection hold-out set, while both methods reach a DSC of 0.08 on the private Radboudumc test set. Compared to our supervised nnU-Net, however, reaching 0.68 on the private Radboudumc set, both methods vastly underperform.

For detection, both proposed methods reached a precision and F1 score close to 0.00, ranging from 0.01 to 0.03 and indicating that there are many false positives. For recall, however, we once again see that our DDPM-based method seems to outperform our DDIM-based method, reaching 0.16 vs. 0.06 on our nnDetection hold-out set and 0.15 vs. 0.03 on the private Radboudumc set. Interestingly, both methods show a relatively high standard deviation in recall on both test sets, indicating that recall varies significantly across different cases. This is especially true for DDPM, which has a standard deviation in recall of 0.30 on the nnDetection hold-out set

Table 6.1: **Segmentation and Detection performance of optimal configurations for all methods on both test sets.** Performance is evaluated at an IoU threshold of 0.2. For each metric, we report the mean value with standard deviation in parentheses. The best performing method for each metric is highlighted in bold. Diffusion-based results were generated with $L = 500$, $s = 1600$ for DDPM and $L = 500$, $s = 1800$ for DDIM.

nnDetection Test	DSC ↑	Precision ↑	Recall ↑	F1-score ↑
DDPM	**0.12 (±0.10)**	0.02 (±0.04)	0.16 (±0.30)	0.03 (±0.06)
DDIM	0.07 (±0.09)	0.01 (±0.09)	0.04 (±0.18)	0.02 (±0.10)
nnDetection	N/A	**0.51 (±0.33)**	**0.85 (±0.26)**	**0.63 (±0.26)**
Radboudumc Test	DSC ↑	Precision ↑	Recall ↑	F1-score ↑
DDPM	0.08 (±0.10)	0.01 (±0.02)	0.15 (±0.28)	0.02 (±0.03)
DDIM	0.08 (±0.11)	0.02 (±0.10)	0.03 (±0.07)	0.02 (±0.05)
nnDetection	N/A	0.51 (±0.28)	**0.73 (±0.28)**	0.55 (±0.24)
nnU-Net	**0.68 (±0.25)**	**0.78 (±0.30)**	0.67 (±0.29)	**0.69 (±0.26)**

and 0.28 on the private Radboudumc set, compared to 0.18 and 0.07 for DDIM respectively.

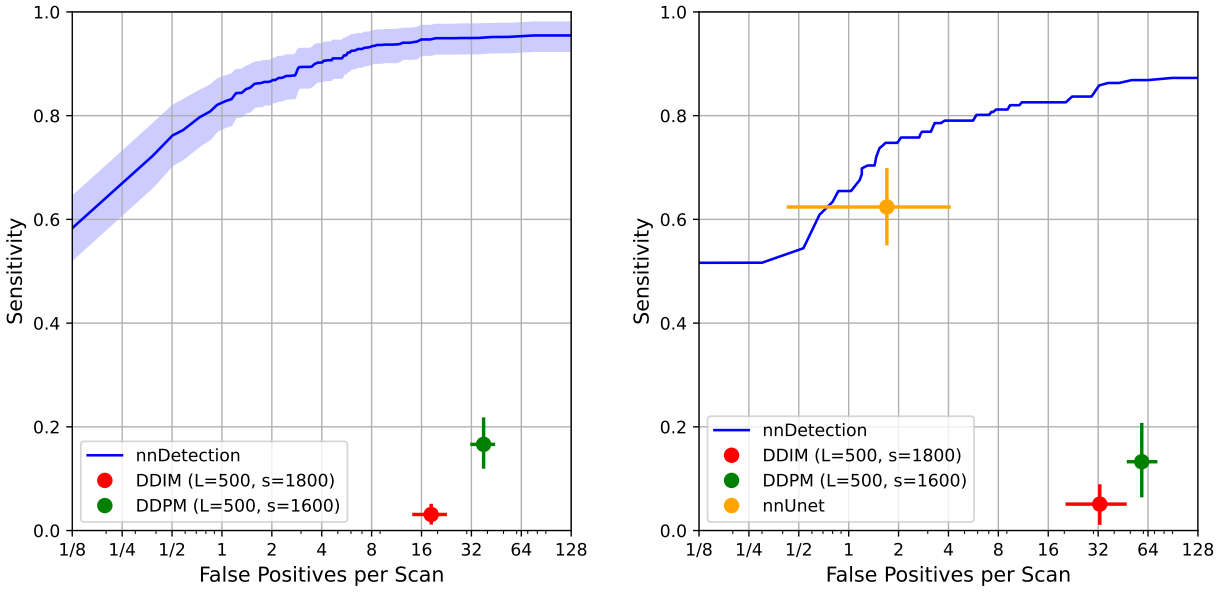
We tested our a priori hypothesis that our DDIM-based method would be superior to our DDPM-based method in terms of median F1 score. We failed to reject the null hypothesis on both the nnDetection hold-out set ($p > 0.999$, adjusted $\alpha = 0.01$) and the private Radboudumc test set ($p = 0.78$, adjusted $\alpha = 0.007$), indicating we could not demonstrate statistical superiority of DDIM over DDPM. The median difference in F1 score was 0.00 for both test sets (mean difference: -0.00 for the nnDetection hold-out set and -0.01 for the private Radboudumc test set).

Both methods are vastly outperformed by our supervised methods, with nnDetection reaching an F1 score of 0.63 on our nnDetection hold-out set and 0.55 on the private Radboudumc set. On the private Radboudumc set, nnU-Net reached an F1 score of 0.69, outperforming both unsupervised methods, as well as nnDetection. We tested whether nnDetection was superior to nnU-Net on the private Radboudumc set in terms of F1 score, but failed to reject the null hypothesis ($p > 0.999$, adjusted $\alpha = 0.008$, with nnU-Net showing higher mean F1 scores (mean difference: -0.16, median difference: -0.06).

Lastly, we tested whether our diffusion-based methods were non-inferior to nnDetection. Both DDIM and DDPM failed to demonstrate non-inferiority on both test sets ($p > 0.999$ for all tests), with substantial median F1 score differences ranging from -0.41 to -0.53, confirming the large performance

gap between unsupervised and supervised approaches.

In the results in Table 6.1, it also seems that nnDetection, evaluated here at a confidence threshold of 0.5, has substantially lower precision than nnU-Net, resulting in a lower F1 score. In Figure 6.1, however, we see that when we evaluate nnDetection at a continuous range of confidence thresholds, resulting in a FROC curve, nnDetection actually seems to outperform nnU-Net at the same number of false positives per scan. This figure plots the sensitivity against the number of false positives per scan for all models, together with 95% confidence intervals calculated by bootstrapping with 1000 bootstraps.



(a) nnDetection Test Set

(b) Radboudumc Private Test Set

Figure 6.1: **Detection performance of our proposed methods compared to the supervised baselines.** The sensitivity is plotted against the average number of false positives per scan for all models on the nnDetection test set (a) and the Radboudumc private test set (b). The nnDetection model was evaluated at a continuous range of confidence intervals between 0 and 1, resulting in a FROC curve. nnU-Net is evaluated by taking the class with the highest confidence for each voxel, resulting in a single point. The diffusion-based models are evaluated at the threshold determines by the Otsu method, similarly resulting in a single point. Bootstrapping was performed with 1000 bootstraps to calculate 95% confidence intervals. These intervals are depicted by the shaded band around the FROC curve and the hairs of the points. Diffusion-based results were generated with $L = 500$, $s = 1600$ for DDPM and $L = 500$, $s = 1800$ for DDIM.

6.1.2 Performance Across Size Ranges

We evaluate the performance of our proposed methods across different lesion sizes, as described in Subsection 5.3.3. Due to limited variability in lesion sizes in the private Radboudumc test set, this analysis is restricted to the nnDetection test set, comparing our methods against the nnDetection baseline.

Table 6.2: **Segmentation and Detection performance of both unsupervised methods and nnDetection on the nnDetection test set for different lesion sizes.** Evaluation was done using only the ground truths that fall within the given size range. Remaining ground truths with a matching prediction (IoU ≥ 0.2) were counted as a true positive, while the ones without were counted as false negatives. Unmatched predictions within the size range were counted as false positives. Metrics were calculated using the same method as described in Table 6.1. We exclude cases with no ground truths within the given size range and report the number of remaining cases within parentheses for each stage. Diffusion-based results were generated with $L = 500$, $s = 1600$ for DDPM and $L = 500$, $s = 1800$ for DDIM.

T stage	Size (cm)	Model	DSC ↑	Precision ↑	Recall ↑	F1-score ↑
T1ai (n = 69)	≤ 2	DDPM	**0.03 (±0.05)**	0.01 (±0.02)	0.14 (±0.29)	0.02 (±0.04)
		DDIM	0.02 (±0.05)	0.01 (±0.03)	0.02 (±0.13)	0.01 (±0.05)
		nnDetection	N/A	**0.18 (±0.24)**	**0.77 (±0.35)**	**0.44 (±0.24)**
T1aii (n = 58)	2−4	DDPM	**0.09 (±0.14)**	0.08 (±0.19)	0.18 (±0.37)	0.10 (±0.22)
		DDIM	0.03 (±0.11)	0.02 (±0.1)	0.05 (±0.22)	0.03 (±0.14)
		nnDetection	N/A	**0.25 (±0.33)**	**0.94 (±0.22)**	**0.62 (±0.26)**
T1b (n = 37)	4−7	DDPM	**0.07 (±0.14)**	0.09 (±0.28)	0.11 (±0.31)	0.09 (±0.28)
		DDIM	0.00 (±0.02)	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.0)
		nnDetection	N/A	**0.26 (±0.39)**	**0.84 (±0.37)**	**0.73 (±0.30)**
T2 (n = 30)	> 7	DDPM	**0.02 (±0.06)**	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)
		DDIM	0.00 (±0.02)	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)
		nnDetection	N/A	**0.43 (±0.43)**	**0.72 (±0.44)**	**0.69 (±0.34)**

Table 6.2 presents the size-stratified performance results. Several key patterns emerge from this analysis:

Both unsupervised methods achieve their highest recall for smaller lesions within the T1a stage ($\leq$ 4cm), with minimal differences between T1ai ($\leq$ 2cm) and T1aii (2−4cm) substages. This size preference is particularly pronounced for DDIM, which demonstrates zero recall for lesions larger than T1a, while DDPM maintains detection capability (recall = 0.11) for T1b lesions (4−7cm). Neither method shows detection capabilities for T2 lesions ($>$ 7cm).

However, both unsupervised methods exhibit poor segmentation quality and precision for the smallest lesions (T1ai stage, $\leq$ 2cm) compared to the T1aii stage (2−4cm). This effect is especially pronounced for DDPM, which achieves a DSC of only 0.03 and precision of 0.01 for T1ai lesions, compared to improved DSC of 0.09 and precision of 0.08 for T1aii lesions.

In comparison, the supervised baseline nnDetection shows a similar performance decline for the smallest lesions (T1ai) but demonstrates substantially superior performance for larger lesions, particularly in the T2 category ($>$ 7cm) where both unsupervised methods fail entirely. Regarding segmentation quality, DDPM achieves its optimal DSC scores for medium-sized lesions (T1aii-T1b: 2−7cm), while DDIM maintains relatively consistent but consistently lower segmentation quality across all detectable size ranges.

6.1.3 Qualitative Analysis

The following qualitative analysis examines representative visual examples to complement the quantitative performance metrics. While this assessment provides valuable insights into model behavior patterns, it should be interpreted in conjunction with the comprehensive quantitative evaluation presented in the preceding sections.

Figure 6.2 presents a visual comparison between the reconstructions, difference maps, and predictions produced by the best-performing DDPM and DDIM models on six representative samples from the nnDetection test set, along with the predictions of nnDetection. These examples were selected to illustrate the range of lesion types and detection challenges encountered. For all predictions, post-processing was applied as described in Section 4.1.1 and Section 5.3.1.

Example a) demonstrates successful detection of a medium-sized (2.9cm) peripheral lesion (on the border of the kidney) by all methods in Table 6.2. Notably, this scan shows clear contrast enhancement and exhibits substantial false positive detections by the diffusion-based methods, particularly visible in the medullary regions.

Example b) presents a challenging case with multiple lesions of varying sizes (0.4−2.4cm. The DDPM model achieves partial detection of several lesions but fails to capture complete lesion boundaries. The DDIM model shows minimal detection capability, while nnDetection also struggles with complete lesion identification in this case. Similar to example a), this scan demonstrates strong contrast enhancement and shows numerous false positive detections by both diffusion-based methods.

Example c) illustrates detection of three smaller ($<$ 2cm) lesions: nnDetection successfully identifies all three, DDPM correctly detects two, and DDIM identifies only one.

Example d) shows another medium sized (3.6cm) peripheral lesion detected by all models, similar to example a) in size and boundary location.

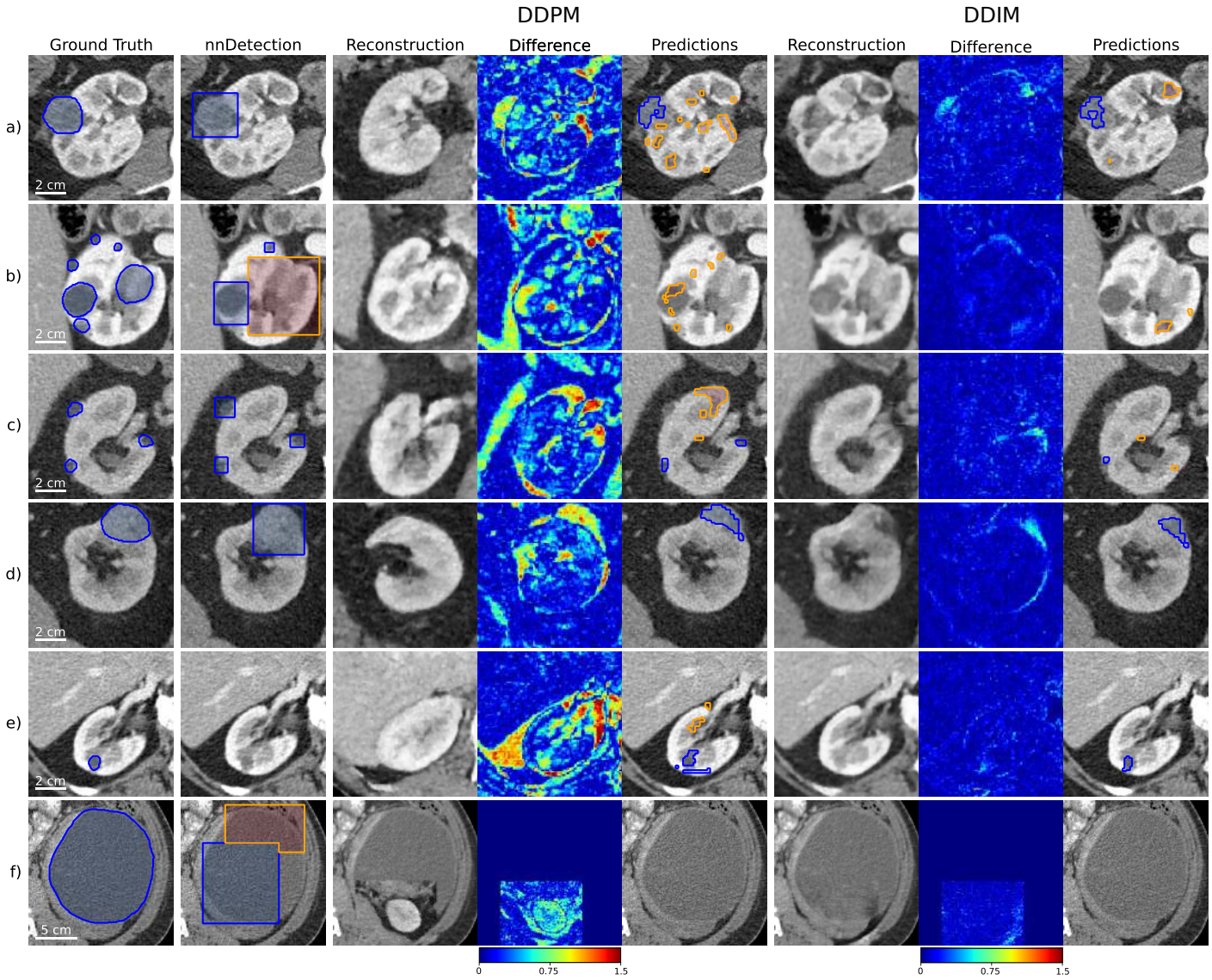


Figure 6.2: **Visual comparison of outputs produced by all models on five cases from the nnDetection test set.** nnU-Net is left out of this comparison, as the nnDetection test set was part of its training data. For nnDetection, the final output bounding box predictions after post-processing are shown. For the diffusion-based models, the reconstruction for each image is shown, along with a difference map between the reconstruction and the original image, as well as the final predictions after post-processing. Reconstructions were generated with $L = 500$, $s = 1600$ for the DDPM and $L = 500$, $s = 1800$ for the DDIM. All examples are shown in the axial plane.

This scan appears to lack contrast enhancement, potentially contributing to the reduced false positive rates observed compared to example a).

Example e) highlights a case where both diffusion-based methods out-

perform the supervised baseline, detecting a small (0.8cm) lesion fully inside the kidney that was overlooked by nnDetection.

Example f) represents a challenging lesion type relatively common in the KiTS dataset, where the lesion substantially exceeds kidney dimensions and falls well outside the region of interest for our diffusion-based methods. All models struggle with accurate detection: nnDetection produces multiple detections for the single lesion, while both diffusion-based methods fail to detect it entirely. Interestingly, the DDPM model reconstructs an entirely new, lesion-free kidney, while DDIM preserves the original anatomy with minimal changes.

Figure 6.3 shows a similar visual comparison for the private Radboudumc test set, including predictions from both supervised baselines (nnDetection and nnU-Net).

Example a) presents a small (0.6cm) lesion correctly detected by nnDetected, but overlooked by all other models. nnU-Net partially detected the lesion, but did not reach a high enough IoU (> 0.2) to count as a detection. Both diffusion-based models show several false positives around the perimeter of the kidney.

Example b) demonstrates a case where nnDetection, nnU-Net, and the DDPM-based method successfully detect the small (0.8cm) lesion, while the DDIM-based method fails to identify it. Our DDPM-based model significantly changed the anatomy of the kidney, leading to substantial activation in the difference map. This was however filtered out correctly by our post-processing.

Example c) shows a small (0.7mm) lesion detected exclusively by nnU-Net, highlighting the superior segmentation and detection capabilities of this method on certain cases.

Examples d) and e) feature larger (4.2 and 6.7cm respectively) lesions extending beyond kidney boundaries. Both supervised methods accurately detect these lesions, with example d) also detected by both unsupervised methods and example e) detected only by the DDPM-based approach.

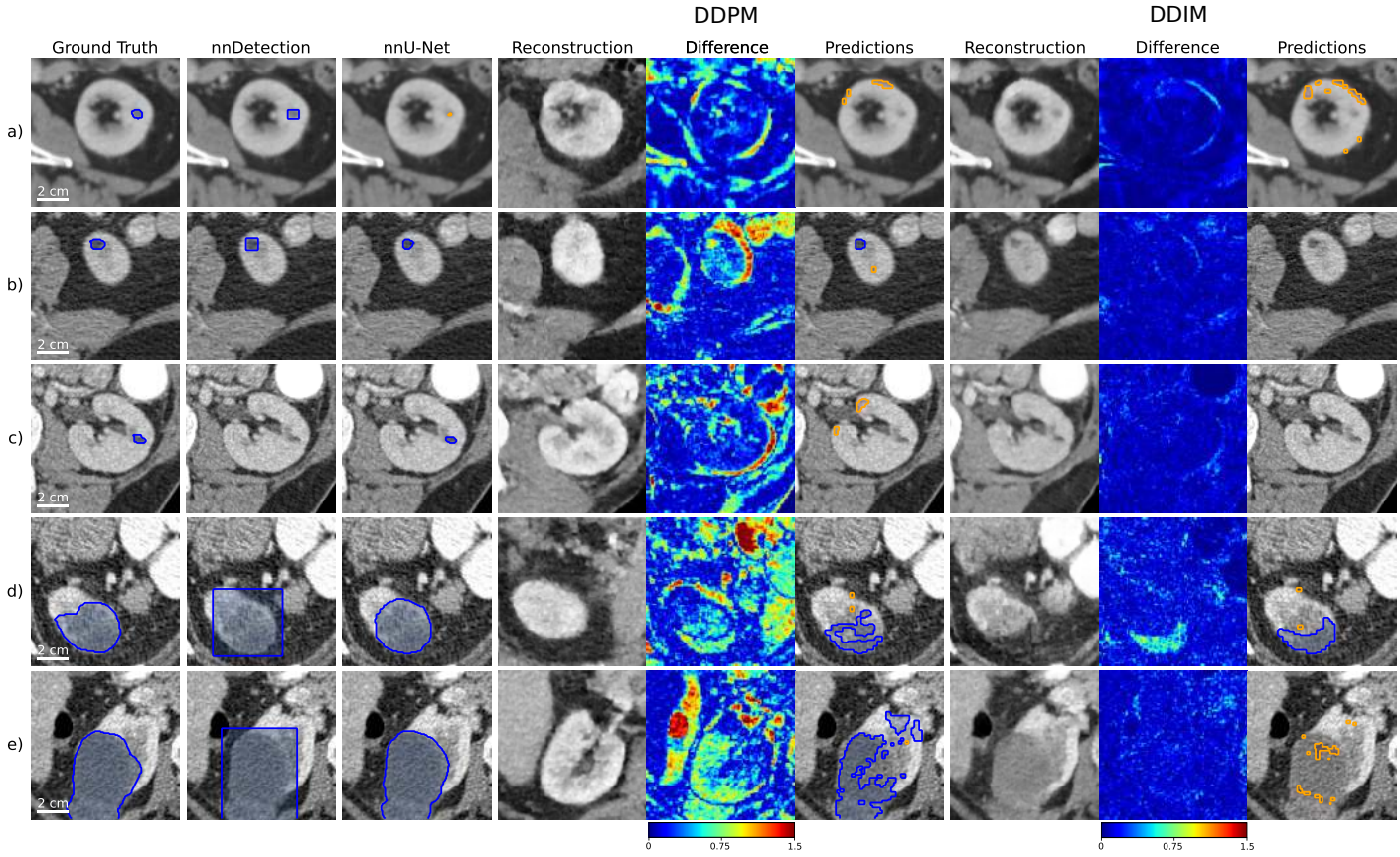


Figure 6.3: **Visual comparison of outputs produced by all models on five cases from the private Radboudumc test set.** For nnDetection, the final output bounding box predictions after post-processing are shown. For nnU-Net, the final output segmentation masks after post-processing are shown. For the diffusion-based models, the reconstruction for each image is shown, along with a difference map between the reconstruction and the original image, as well as the final predictions after post-processing.Reconstructions were generated with $L = 500$, $s = 1600$ for the DDPM and $L = 500$, $s = 1800$ for the DDIM. All examples are shown in the axial plane.

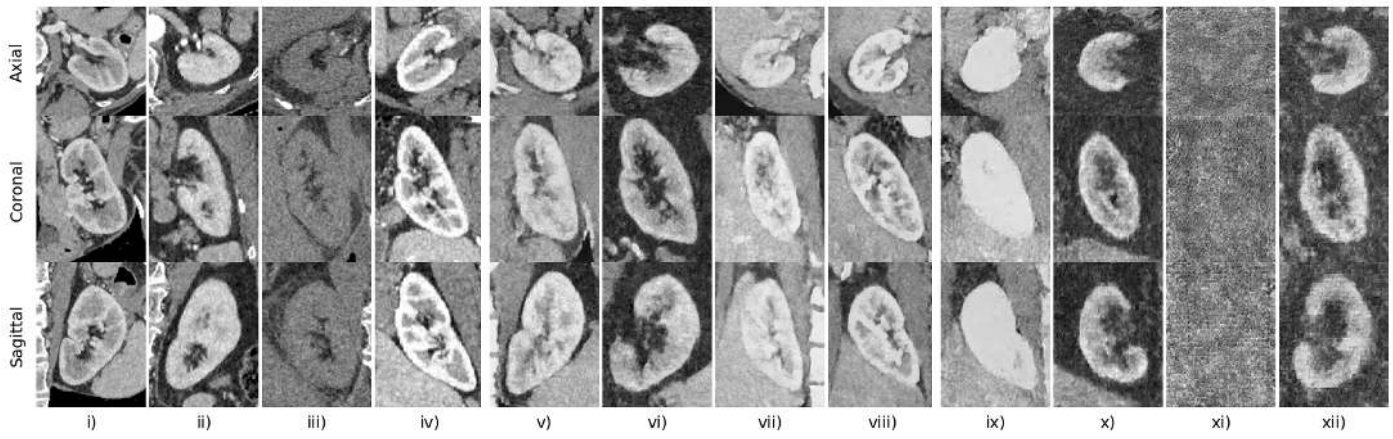
6.2 Performance of Sub-Components

6.2.1 Latent Diffusion Model Performance

Figure 6.4 shows four curated samples showcasing the generative capability of our latent diffusion model using both samplers (DDPM and DDIM). Four samples from real kidney CT scans are included for comparison. More randomly selected samples for each category are included in Appendix B.

The examples in Figure B.1b demonstrate that the diffusion model is capable of generating a variety of kidney CT patches, including both left and right kidneys. The generated samples exhibit variety in contrast phases as well as diversity in the surrounding anatomy. Since the samples generated by the model were not evaluated by expert radiologists, no conclusions can be drawn about the anatomical correctness of the generated kidneys or surrounding anatomy. From a non-expert perspective, however, the overall shape of the kidneys appears anatomically accurate. The model demonstrates limitations in accurately generating the medulla of the kidneys, which is substantially less well-defined in the generated samples compared to the real samples. The hilium and urine collection system are also poorly defined in the generated samples, with the model occasionally generating blurry or malformed regions where the ureter should be located.

The model exhibits notable difficulties when generating completely new samples using DDIM sampling. The generated samples are often of inferior visual quality compared to those generated using DDPM sampling. Occasionally, the model completely fails to properly produce an image, resulting in noisy output as observed in example xii. This issue appears exclusively when generating samples from pure noise and has not occurred in other experiments where the model added and removed noise from existing images.



(a) Real kidney CT patches

(b) DDPM generated kidney CT patches

(c) DDIM generated kidney CT patches

Figure 6.4: **Generation examples of our diffusion model compared to real kidney CT patches.** Cases i-iv show real kidney CT patches, v-viii show kidney CT patches generated by our DDPM model, and ix-xii show kidney CT patches generated by our DDIM model.

We also quantitatively measured the diversity of the generated samples compared to the real kidney CT patches using the MS-SSIM. These results are summarized in Table 6.3. We find that DDPM and DDIM generated samples have similar diversity, reaching an MS-SSIM of 0.073 and 0.074 respec-

Table 6.3: **MS-SSIM values for real and generated kidney CT patches.** We report the mean MS-SSIM value along with the standard deviation in parentheses. A lower MS-SSIM indicates more structural diversity between samples.

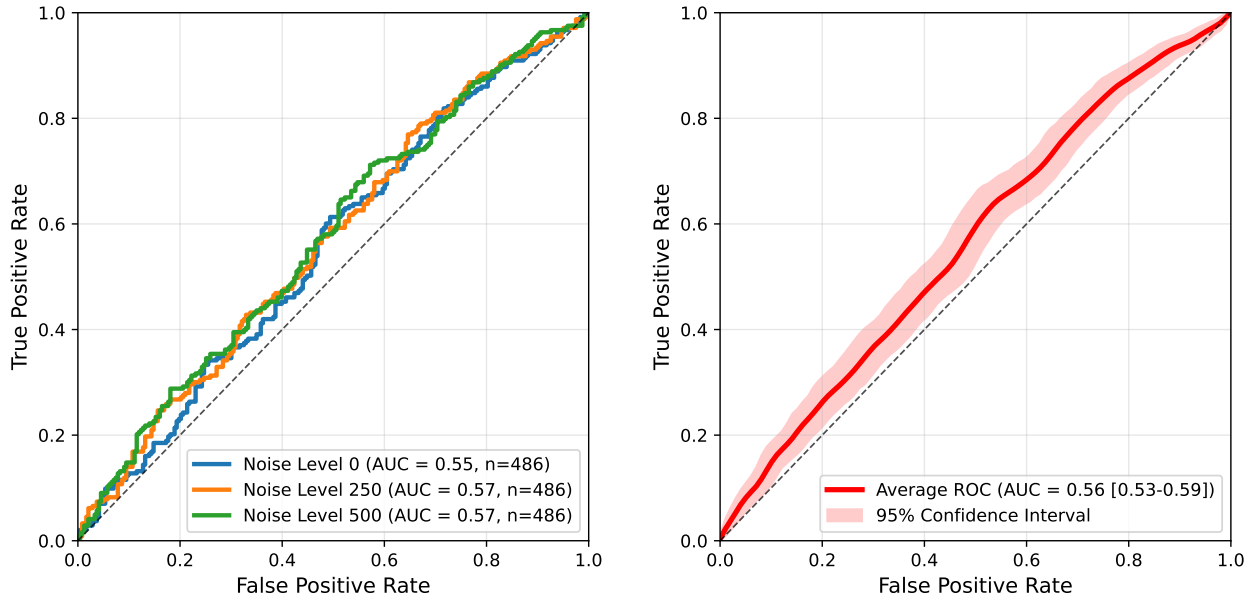
Model	**MS-SSIM ↓**
Real	0.057 (±0.054)
DDPM	0.073 (±0.080)
DDIM	0.074 (±0.080)

tively. Both methods show slightly less diversity than the real samples, which reached an MS-SSIM of 0.057. The standard deviation is relatively high for all three categories (0.54−0.080), indicating a wide range of diversity between individual samples.

6.2.2 Classifier Performance

We evaluated the performance of our stand-alone classifier as described in Subsection 5.4.2. Figure 6.5a shows the ROC curves describing the performance of our classifier on the same dataset at noise levels 0, 250, and 500. We received a similar performance on all three noise levels, reaching an AUC of 0.55, 0.57, and 0.57 for noise levels 0, 250, and 500 respectively.

Looking at the averaged result, we reach an AUC of 0.56 (95% confidence interval: 0.53-0.59). While this is technically above chance performance (AUC=0.500), the difference is minimal and indicates that the classifier struggles substantially to distinguish healthy kidneys from pathological ones.



(a) ROC curve per noise level

(b) ROC curve averaged over noise levels, with 95% confidence interval

Figure 6.5: **Classifier performance on combined Radboudumc test set.** We show the ROC curve and corresponding AUC value for noise level 0, 250 and 500 (a). We also show the ROC curve averaged over these noise levels with 95% confidence intervals calculated by bootstrapping with 1000 bootstraps (b).

6.2.3 TotalSegmentator Performance

We evaluated the performance of TotalSegmentator on both test sets. The results are shown in Table 6.4. TotalSegmentator achieved near-perfect recall (0.99) on the nnDetection test set and perfect recall (1.00) on the Radboudumc test set. Precision was slightly lower, reaching 0.96 on the nnDetection test set and 0.95 on the Radboudumc test set. Overall, TotalSegmentator demonstrated excellent performance for kidney detection across both datasets.

Table 6.4: **Kidney detection performance of TotalSegmentator on our two test sets.** We report the precision, recall, F1 score, and accuracy along with the 95% confidence interval for each value. A kidney was counted as detected if a connected component in the ground truth kidney segmentation map and a left or right kidney segmentation produced by TotalSegmentator have IoU ¿ 0.5. A predicted segmentation is counted as false positive if no connected component can be found in the ground truth map with IoU ¿ 0.5. We excluded one case where the kidneys in the ground truth segmentation were merged into a single connected component.

nnDetection Test	Precision ↑	Recall ↑	F1 Score ↑	Accuracy ↑
Left	0.94 (0.90 - 0.98)	0.99 (0.98 - 1.00)	0.97 (0.94 - 0.99)	0.94 (0.89 - 0.97)
Right	0.97 (0.94 - 0.99)	0.99 (0.98 - 1.00)	0.98 (0.96 - 1.00)	0.96 (0.92 - 0.99)
Both	0.96 (0.93 - 0.98)	0.99 (0.98 - 1.00)	0.97 (0.96 - 0.99)	0.95 (0.92 - 0.97)
Radboudumc Test	Precision ↑	Recall ↑	F1 Score ↑	Accuracy ↑
Left	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)
Right	0.89 (0.77 - 1.00)	1.00 (1.00 - 1.00)	0.94 (0.87 - 1.00)	0.90 (0.77 - 1.00)
Both	0.95 (0.88 - 1.00)	1.00 (1.00 - 1.00)	0.97 (0.94 - 1.00)	0.95 (0.88 - 1.00)

Chapter 7

Discussion

In this chapter, we discuss the answers to our research questions, thereby highlighting our key findings. We then interpret our main findings, combining context from our work and existing literature. Finally, we examine the limitations of our approach alongside directions for future research.

7.1 Answering Our Research Questions

7.1.1 Weakly Supervised Method Performance

Table 6.1 indicates that DDPM-based reconstruction outperforms DDIM-based reconstruction on anomaly segmentation, achieving DSC scores of 0.12 vs. 0.07 on the nnDetection test set. However, given the lack of statistical testing, the relatively small effect size, and the equivalent DSC scores of 0.08 on the Radboudumc test set, we cannot draw definitive conclusions about their relative segmentation performance, thus limiting our ability to answer **RQ1a** with certainty.

For detection performance, Table 6.1 shows minimal differences between DDPM-based and DDIM-based methods in precision (0.01–0.02) and F1 score (0.02–0.03). However, substantial differences emerge in recall performance: the DDPM-based method achieved recall values of 0.16 and 0.15 on the nnDetection and Radboudumc test sets, respectively, while the DDIM-based method achieved only 0.04 and 0.03. Although we did not perform statistical significance testing on recall specifically, the large effect size across both datasets suggests superior recall for the DDPM-based method. This is supported by Figure 6.2, where the 95% confidence intervals demonstrate the DDPM-based method's superior recall on the nnDetection test set. However, the DDPM-based method generates substantially more false positives than the DDIM-based method on both test sets. We conclude that the DDPM-based method outperforms the DDIM-based method in recall at the

expense of higher false positive rates, answering **RQ1b**.

Our size-stratified analysis revealed distinct performance patterns across lesion sizes for both diffusion-based methods. Both methods achieved their optimal performance on smaller lesions, with performance declining dramatically for larger lesions. Specifically, the DDIM-based method demonstrated complete failure for lesions larger than T1a, while the DDPM-based method maintained detection capability for T1b lesions (4−7cm) but failed entirely on T2 lesions ($>$ 7cm).

Within the smaller lesion categories, we observed notable quality differences. Our DDPM-based method exhibited substantially lower DSC and precision for T1ai lesions ($\leq$ 2cm) compared to T1aii lesions (2−4cm), which we attribute to false positives resulting from slight anatomical deviations in reconstructions. Beyond this specific pattern, the performance differences within stage T1a for the DDIM-based method and within stage T1 for the DDPM-based method are relatively modest, preventing us from drawing definitive conclusions about their relative performance within these subgroups.

Overall, we can conclude that both methods face significant challenges with larger lesions: the DDPM-based method shows a steep performance decline for T2 lesions, while the DDIM-based method fails entirely on both T1b and T2 lesions. Additionally, the DDPM-based method generates substantially more false positives in the T1ai range compared to other size categories. This comprehensive size-dependent analysis answers **RQ1c**.

7.1.2 Supervised Method Comparison

Addressing **RQ2a**, nnDetection achieved higher recall at equivalent false positive rates (as shown in Figure 6.2), while nnU-Net demonstrated superior overall F1 performance (0.69 vs. 0.55) on the Radboudumc test set. Although statistical testing failed to demonstrate nnDetection's superiority over nnU-Net, the substantial effect size in favor of nnU-Net suggests superior detection performance for nnU-Net in terms of F1 score, while nnDetection may excel in recall at equivalent false positive rates.

For segmentation performance, nnU-Net achieved substantially higher DSC scores (0.68) compared to both diffusion-based methods (0.08), demonstrating that our proposed methods failed to achieve competitive segmentation performance relative to supervised methods, answering **RQ2b**.

Our proposed methods demonstrated inferior detection performance compared to supervised baselines. As shown in Table 6.1, the precision and F1 scores of both diffusion-based methods are at least an order of magnitude lower than those achieved by nnDetection and nnU-Net on both

datasets. In terms of recall, our best performing method (DDPM-based method) achieved 0.16 on the nnDetection test set compared to 0.85 for nnDetection, and 0.15 on the Radboudumc test set compared to 0.73 and 0.67 for nnDetection and nnU-Net, respectively. While we cannot conclude inferiority solely from our failure to statistically demonstrate non-inferiority, the substantially lower performance across all metrics clearly demonstrates that our weakly supervised methods cannot achieve detection performance comparable to supervised methods, answering **RQ2c**.

7.1.3 Sub-Component Analysis

Our qualitative evaluation of generation using latent diffusion reveals that DDPM-based sampling produces good visual fidelity and reasonable anatomical accuracy in generated output. While the model successfully captures kidney outlines, it struggles with fine internal structures including the hilum, medulla, and urine collection system. DDIM-based sampling, however, produced unpredictable generation with reduced anatomical accuracy and occasional complete failures.

Quantitative diversity analysis shows higher MS-SSIM scores for both sampling methods (0.073–0.074) compared to real images (0.057), indicating somewhat reduced structural diversity between generated samples. However, the difference is relatively small (0.016) and likely reflects reduced variability in imaging characteristics such as contrast patterns and intensity distributions.

We conclude that our diffusion model generates anatomically diverse and reasonably accurate kidney images, with DDPM-based sampling being substantially more reliable than DDIM-based sampling for generating new examples, answering **RQ3a**.

For **RQ3b**, our classifier achieved an AUC of only 0.56 (95% CI: 0.53–0.59), barely exceeding chance performance. This poor performance persisted even when evaluated on clean images (noise level 0), indicating fundamental challenges in distinguishing healthy from pathological tissue in the VQ-GAN latent space.

Regarding **RQ3c**, TotalSegmentator demonstrated excellent kidney detection performance across both test sets, achieving near-perfect recall (0.99 on nnDetection test set, 1.00 on Radboudumc test set) and high precision (0.96 and 0.95, respectively).

7.2 Interpretation of Findings

7.2.1 Low Segmentation and Detection Performance

Despite promising results from previous work on diffusion-based anomaly detection [67], our diffusion-based models achieved substantially lower performance compared to supervised methods on both detection and segmentation tasks. With precision values of 0.01–0.02 across both test sets, the vast majority of detected anomalies were false positives rather than actual lesions. While weakly supervised methods are expected to underperform compared to supervised methods trained on sufficient data, the magnitude of this performance gap requires explanation.

This substantial performance difference can be attributed to several factors. First, the anatomical complexity of abdominal imaging presents significant challenges for reconstruction-based anomaly detection. The abdominal region features complex internal kidney structure with cortex and medulla, heterogeneous surrounding anatomy including fat, vessels, and other organs, and variable contrast enhancement patterns. This complexity is supported by previous work demonstrating that the same 3D latent diffusion architecture achieved substantially higher generation quality on brain MRI compared to other anatomical regions [35]. Similarly, existing work on 2D pancreas tumor segmentation using diffusion-based anomaly detection has shown lower segmentation performance compared to similar work on brain MRI [2], highlighting the difficulties posed by abdominal anatomy. The high variability in kidney anatomy and contrast enhancement patterns leads to numerous false positives, particularly in medullary regions where our model struggled to accurately reconstruct normal enhancement patterns.

Second, our chosen architecture has specific limitations that impact anomaly detection performance. Most critically, our classifier achieved an AUC of only 0.56, barely exceeding chance performance. This poor performance may be partially attributed to our use of latent diffusion, which requires the classifier to perform classification in a quantized latent space rather than pixel space, likely making it more challenging to distinguish healthy from pathological tissue. The choice of VQ-GAN as the autoencoder component may have further exacerbated reconstruction challenges, as the adversarial and perceptual losses used to reduce blurriness may introduce noise when computing voxel-wise differences between original and reconstructed images.

Finally, reconstruction-based anomaly detection faces two fundamental challenges that contribute to poor performance. First, the method is inherently dependent on intensity differences between reconstructed and original images, meaning that lesions with similar intensity characteristics to surrounding healthy tissue—particularly cystic lesions or those with low contrast enhancement—remain undetected. Second, the approach gener-

ates numerous false positives from anatomical differences and variations in contrast enhancement between reconstructed and original images. These anatomical inconsistencies are particularly pronounced around the perimeter of the kidney, and the kidney hilum. This false positive problem is particularly pronounced for our DDPM-based method, which is consistent with literature [67].

7.2.2 DDPM-Based Method Superiority Over DDIM-Based Method

Our finding that the DDPM-based method outperformed the DDIM-based method contradicts existing literature on diffusion-based anomaly detection [67]. We attribute this effect primarily to the near-chance performance (AUC: 0.56) of our classifier guidance mechanism. Although the denoising U-Net is trained on only healthy images, the DDIM-based method's deterministic nature makes it more dependent on the quality of the guidance provided by the classifier than the DDPM-based method, as evident from the mathematical formulation where DDIM's forward and reverse processes are exact inverses (Equation 4.2 and Equation 4.3). This effect is less pronounced for the DDPM-based method, which relies on noise being introduced during sampling (Equation 2.8).

Another potential contribution to this gap in performance could be the unstable generation when using DDIM-based sampling. However, while this effect was very pronounced when generating new samples from noise, it was not observed during reconstruction of real images. This leads us to conclude that the effect only occurs when sampling from timestep 0, and likely did not impact the detection and segmentation performance of our DDIM-based reconstruction method.

7.2.3 Poor Performance on Large Lesions

Our size-stratified analysis revealed that both diffusion-based methods performed best on smaller lesions within the T1a category ($\leq$ 4cm), but showed different patterns of performance decline for larger lesions. The DDIM-based method demonstrated complete failure for lesions larger than T1a, showing dramatic performance decline immediately above this threshold. In contrast, the DDPM-based method maintained detection capability for T1b lesions ($4-7$cm) but failed entirely on T2 lesions ($>$ 7cm).

This size-dependent performance pattern can be attributed to several factors. First, larger lesions often extend beyond kidney boundaries or substantially alter kidney anatomy, making it difficult for a model trained on healthy kidney patches to reconstruct appropriate "healthy" versions. Second, our region of interest size of $96 \times 96 \times 128$mm cannot contain all lesions in stage T2 ($>$ 70mm). As demonstrated in our qualitative analysis, our

model is unable to reconstruct lesions that fall partially outside this region of interest. Finally, TotalSegmentator has been shown to have reduced segmentation accuracy for kidneys with pathologies [10]. This could lead to segmentation maps that exclude portions of the kidney and lesion, resulting in incorrect masking during post-processing.

7.3 Limitations and Future Directions

7.3.1 Classifier Performance Bottleneck

Additionally, our classifier's inadequate performance likely severely limited the effectiveness of our guidance mechanism, particularly for the DDIM-based method, which relies entirely on classifier gradients for reconstruction differences. This bottleneck effectively undermined the core principle of guided reconstruction that underpins diffusion-based anomaly detection, explaining both the overall poor performance and the unexpected superiority of the DDPM-based method over the DDIM-based method.

Future work should focus on improving classifier performance through several approaches: (1) using architectures that do not rely on latent diffusion, such as 3D wavelet diffusion [21], allowing classifiers to work directly in pixel space; (2) exploring improved pseudo-label generation through sophisticated natural language processing of radiology reports using large language models [1] or expert validation; and (3) investigating alternative guidance techniques such as classifier-free guidance [12] or implicit guidance [7].

7.3.2 High False Positive Rates

The high number of false positives significantly contributes to our methods' low performance. While our DDIM-based method eliminated many false positives through more accurate anatomical reconstruction, our DDPM-based method generated numerous false positives that persisted after post-processing. Although morphological operations and connected component analysis helped reduce anatomical inconsistencies and noise, these post-processing steps come at the cost of potentially removing smaller lesions, creating a trade-off between false positive reduction and sensitivity to small pathologies.

Future work should address false positives through several technical approaches: (1) replacing Gaussian sampling with Simplex sampling as introduced in AnoDDPM [68]; (2) implementing iterative masking, stitching, and resampling as in AutoDDPM [6]; and (3) developing dedicated false positive reduction networks, though current implementations have relied on supervised methods [27].

7.3.3 Limited Dataset Scope and Evaluation Constraints

Our evaluation was constrained by limited population diversity, with data from only two distinct populations: Netherlands patients treated at Radboudumc and US patients from the KiTS dataset. While KiTS imaging originates from multiple institutions, it represents a US-only population, limiting generalizability across different healthcare systems and patient demographics. Additionally, the size distribution across both datasets (95% T1a lesions in the Radboudumc test set, 80% in the nnDetection test set) severely constrained comprehensive size-stratified evaluation.

Future work should prioritize developing standardized public benchmarks that facilitate fair evaluation of both supervised and unsupervised methods across varied imaging protocols, patient populations, and lesion size distributions. Multi-institutional collaboration across different geographic regions would be particularly valuable for establishing generalizability across diverse clinical scenarios.

7.3.4 Limited Quantitative Evaluation of Generation Quality

We were unable to perform comprehensive quantitative analysis of generation fidelity. While metrics like Fréchet Inception Distance (FID) [28] are commonly used to evaluate generation quality by measuring distances between feature distributions of generated and real samples, meaningful interpretation requires comparison to suitable baselines on the same dataset.

7.3.5 Inherent Dependencies on Supervised Components

While our method is primarily based on weakly supervised learning, it retains dependencies on supervised components inherent to the approach. Our method required annotated data for hyperparameter optimization to determine optimal noise levels and guidance strength, with Figure 5.1 showing high sensitivity to these parameters, particularly for the DDIM-based method. Additionally, our pipeline relies on TotalSegmentator for kidney segmentation, which was trained in a supervised manner.

These limitations highlight the practical constraints of achieving truly unsupervised medical image analysis. While completely eliminating the need for annotated data is not feasible, future work could focus on reducing required data through more efficient hyperparameter optimization methods and exploring full-scan anomaly detection approaches, though this would introduce substantial computational complexity and likely increase false positives outside target organs.

Chapter 8

Conclusions

In this thesis, we explored the use of latent diffusion models for weakly supervised kidney anomaly detection on fully volumetric contrast-enhanced abdominal CT imaging. To the best of our knowledge, this represents the first fully 3D diffusion-based anomaly detection pipeline for abdominal CT. We introduced a novel pipeline using TotalSegmentator, 3D latent diffusion, classifier-guided reconstruction, and post-processing techniques. We evaluated segmentation and detection performance for two different diffusion sampling methods, DDPM and DDIM. In contrast to previous work, we also evaluated our weakly supervised methods against two supervised baselines: a pre-trained nnU-Net model and a newly trained nnDetection model.

We found that both our weakly supervised methods underperformed compared to the supervised baselines. For segmentation performance, our best-performing diffusion-based method achieved a DSC of only $0.08-0.12$, compared to supervised baselines achieving DSCs of $0.51-0.68$. For detection performance, our best-performing diffusion-based method achieved a precision of only $0.01-0.02$ and recall of $0.15-0.16$, compared to supervised baselines achieving precision of $0.51-0.78$ and recall of $0.67-0.85$, measured at an IoU threshold of 0.2. Our size-stratified analysis revealed that both methods performed best on smaller lesions ($\leq$ 7 cm for our DDPM-based method and $\leq$ 4 cm for our DDIM-based one) with substantial performance decline for larger lesions.

Qualitative analysis of the predictions produced by our proposed methods revealed that lesions on the boundary of the kidney and lesions that have high contrast to the surrounding tissue are detected more often, while lesions that present similar intensity to surrounding tissue are often segmented only partially or overlooked completely. Additionally, our analysis showed that many false positives were introduced during the reconstruction process, particularly on the boundaries of the kidney or in the medulla.

Counter to our expectations based on existing literature, our results

showed that our DDPM-based method outperformed our DDIM-based method in terms of segmentation performance and recall, while our DDIM-based method showed slightly fewer false positives. We attribute this performance gap to the poor performance of our classifier (AUC = 0.56). Due to the deterministic nature of DDIM sampling, this method is more dependent on classifier guidance than our DDPM-based method.

Overall, our findings highlight the potential of latent diffusion models for weakly supervised anomaly detection, while also underscoring the significant challenges that remain in improving their performance to match supervised methods, particularly in complex anatomical regions and real-world clinical scenarios. Our comprehensive evaluation provides valuable insights for future improvements in this promising research direction.

Bibliography

[1] Al Mohamad, F., Donle, L., Dorfner, F., Romanescu, L., Drechsler, K., Wattjes, M.P., Nawabi, J., Makowski, M.R., Häntze, H., Adams, L., et al.: Open-source large language models can generate labels from radiology reports for training convolutional neural networks. Academic Radiology **32**(5), 2402–2410 (2025)

[2] Babaei, R., Cheng, S., Thai, T., Zhao, S.: Pancreatic tumor segmentation as anomaly detection in CT images using denoising diffusion models. arXiv preprint arXiv:2406.02653 (2024)

[3] Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9252–9260 (2018)

[4] Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nnDetection: a self-configuring method for medical object detection. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 530–539. Springer (2021)

[5] Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T., Boss, A.: Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investigative radiology **52**(7), 434–440 (2017)

[6] Bercea, C.I., Neumayr, M., Rueckert, D., Schnabel, J.A.: Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. arXiv preprint arXiv:2305.19643 (2023)

[7] Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion models with implicit guidance for medical anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 211–220. Springer (2024)

[8] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)

[9] Blausen.com staff: Medical gallery of blausen medical 2014. WikiJournal of Medicine **1**(2) (2014). https://doi.org/10.15347/wjm/2014.010

[10] de Boer, S., Häntze, H., Venkadesh, K.V., Buser, M.A., Mamani, G.E.H., Xu, L., et al.: Robust kidney abnormality segmentation: A validation study of an AI-based framework. arXiv preprint arXiv:2505.07573 (2025)

[11] Capitanio, U., Montorsi, F.: Renal cancer. The Lancet **387**, 894–906 (2016). https://doi.org/10.1016/S0140-6736(15)00046-X

[12] Che, Y., Rafsani, F., Shah, J., Siddiquee, M.M.R., Wu, T.: AnoFPDM: Anomaly detection with forward process of diffusion models for brain MRI. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1113–1122 (2025)

[13] Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)

[14] Cirillo, L., Innocenti, S., Becherucci, F.: Global epidemiology of kidney cancer. Nephrology Dialysis Transplantation **39**(6) (2024). https://doi.org/10.1093/ndt/gfae036

[15] Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)

[16] Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with Gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016)

[17] ud din, N.M., Dar, R.A., Rasool, M., Assad, A.: Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Computers in Biology and Medicine **149**, 106073 (2022). https://doi.org/10.1016/j.compbiomed.2022.106073

[18] Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)

[19] Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Deep Learning for Medical Anomaly Detection – A Survey. ACM Comput. Surv. **54**(7), 141:1–141:37 (2021). https://doi.org/10.1145/3464423

[20] Fontanella, A., Mair, G., Wardlaw, J., Trucco, E., Storkey, A.: Diffusion models for counterfactual generation and anomaly detection in brain images. IEEE Transactions on Medical Imaging (2024)

[21] Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: WDM: 3D wavelet diffusion models for high-resolution medical image synthesis. In: MICCAI workshop on deep generative models. pp. 11–21. Springer (2024)

[22] Goetz, L., Seedat, N., Vandersluis, R., van der Schaar, M.: Generalization—a key challenge for responsible AI in patient-facing clinical applications. NPJ Digital Medicine **7**(1), 126 (2024)

[23] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Advances in Neural Information Processing Systems **27** (2014)

[24] Graham, M.S., Pinaya, W.H.L., Wright, P., Tudosiu, P.D., Mah, Y.H., Teo, J.T., Jäger, H.R., Werring, D., Nachev, P., Ourselin, S., et al.: Unsupervised 3D out-of-distribution detection with latent diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 446–456. Springer (2023)

[25] Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT (2023)

[26] Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)

[27] Hendrix, W., Hendrix, N., Scholten, E.T., Mourits, M., Trap-de Jong, J., Schalekamp, S., Korst, M., Van Leuken, M., Van Ginneken, B., Prokop, M., et al.: Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Communications medicine **3**(1), 156 (2023)

[28] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)

[29] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

[30] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pp. 65–70 (1979)

[31] Hu, X., Jin, C.: AnoDODE: Anomaly detection with diffusion ODE. arXiv preprint arXiv:2310.06420 (2023)

[32] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[33] Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in CT. European Radiology **29**(3), 1391–1399 (Mar 2019). https://doi.org/10.1007/s00330-018-5695-5

[34] Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis **88**, 102846 (2023). https://doi.org/10.1016/j.media.2023.102846

[35] Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., et al.: Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports **13**(1), 7303 (2023)

[36] Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)

[37] Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)

[38] Lång, K., Josefsson, V., Larsson, A.M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., Rosso, A.: Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. The Lancet Oncology **24**(8), 936–944 (2023)

[39] Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health **1**(6), e271–e297 (2019)

[40] Mamani, G.E.H., Lessmann, N., Scholten, E.T., Prokop, M., Jacobs, C., van Ginneken, B.: Kidney abnormality segmentation in thorax-abdomen CT scans. arXiv preprint arXiv:2309.03383 (2023)

[41] Mazandarani, F.N., Babyn, P., Alirezaie, J.: UNeXt: a low-dose CT denoising UNet model with the modified ConvNeXt block. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

[42] Medicine, H.: Kidney cancer staging. https://www.hopkinsmedicine.org/health/conditions-and-diseases/kidney-cancer-staging, accessed: 2025-08-13

[43] Meyer, C.R., Johnson, T.D., McLennan, G., Aberle, D.R., Kazerooni, E.A., MacMahon, H., et al.: Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods. Academic Radiology **13**(10), 1254–1265 (Oct 2006). https://doi.org/10.1016/j.acra.2006.07.012

[44] Mykula, H., Gasser, L., Lobmaier, S., Schnabel, J.A., Zimmer, V., Bercea, C.I.: Diffusion models for unsupervised anomaly detection in fetal brain ultrasound. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 220–230. Springer (2024)

[45] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076

[46] Pinaya, W.H., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., et al.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 705–714. Springer (2022)

[47] Plana, D., Shung, D.L., Grimshaw, A.A., Saraf, A., Sung, J.J., Kann, B.H.: Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Network Open **5**(9), e2233946–e2233946 (2022)

[48] Ragab, M.G., Abdulkadir, S.J., Muneer, A., Alqushaibi, A., Sumiea, E.H., Qureshi, R., et al.: A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). IEEE Access **12**, 57815–57836. https://doi.org/10.1109/ACCESS.2024.3386826

[49] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788. IEEE (2016). https://doi.org/10.1109/CVPR.2016.91

[50] Rinze Reinhard, M.v.d.Z.C., Smithuis, R.: The radiology assistant: Solid renal masses. https://radiologyassistant.nl/abdomen/kidney/solid-masses, accessed: 2025-08-13

[51] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

[52] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

[53] Saha, A., Bosma, J.S., Twilt, J.J., van Ginneken, B., Bjartell, A., Padhani, A.R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., et al.: Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. The Lancet Oncology **25**(7), 879–887 (2024)

[54] Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis **54**, 30–44 (2019)

[55] Smithuis, R.: The radiology assistant: CT contrast injection and protocols. https://radiologyassistant.nl/more/ct-protocols/ct-contrast-injection-and-protocols, accessed: 2025-06-27

[56] Sobek, J., Medina Inojosa, J.R., Medina Inojosa, B.J., Rassoulinejad-Mousavi, S., Conte, G.M., Lopez-Jimenez, F., Erickson, B.J.: MedY-

OLO: a medical image object detection framework. Journal of Imaging Informatics in Medicine pp. 1–9 (2024)

[57] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/sohl-dickstein15.html

[58] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[59] Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V., Waddell, N.: Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine **13**(1), 152 (2021)

[60] UK, C.R.: TNM stages of kidney cancer. https://www.cancerresearchuk.org/about-cancer/kidney-cancer/stages-types-grades/tnm, accessed: 2025-08-13

[61] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in Neural Information Processing Systems **30** (2017)

[62] Venkadesh, K.V., Setio, A.A., Schreuder, A., Scholten, E.T., Chung, K., W. Wille, M.M., Saghir, Z., van Ginneken, B., Prokop, M., Jacobs, C.: Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. Radiology **300**(2), 438–447 (2021)

[63] Wang, P., Berzin, T.M., Brown, J.R.G., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al.: Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut **68**(10), 1813–1819 (2019)

[64] Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)

[65] Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)

[66] Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics bulletin **1**(6), 80–83 (1945)

[67] Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: International Conference on Medical image computing and computer-assisted intervention. pp. 35–45. Springer (2022)

[68] Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 650–656 (2022)

[69] Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**(1), 60–66 (2019)

[70] Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E.I.C., Xu, Y.: MRI cross-modality image-to-image translation. Scientific Reports **10**(1), 3753 (2020)

[71] Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941 (2018)

Appendix A

Training Hyperparameters

Table A.1: VQ-GAN Hyperparameters

Parameter	**Value**
Input size	[128, 96, 96]
Downsampling factor	[4,4,4]
Embedding dimensions	8
Codebook size	16,384
Precision	32 bit
Optimizer	Adam
Learning rate	0.0003
β_1, β_2	0.5, 0.9
Loss weights	
Volume Discriminator	1.0
Slice Discriminator	1.0
Perceptual	4.0
Reconstruction (L1)	4.0
Discriminator start	After 10,000 iterations
Batch size	4
Gradient accumulation	Every 2 epochs
Maximum iterations	100,000

Table A.2: DDPM Hyperparameters

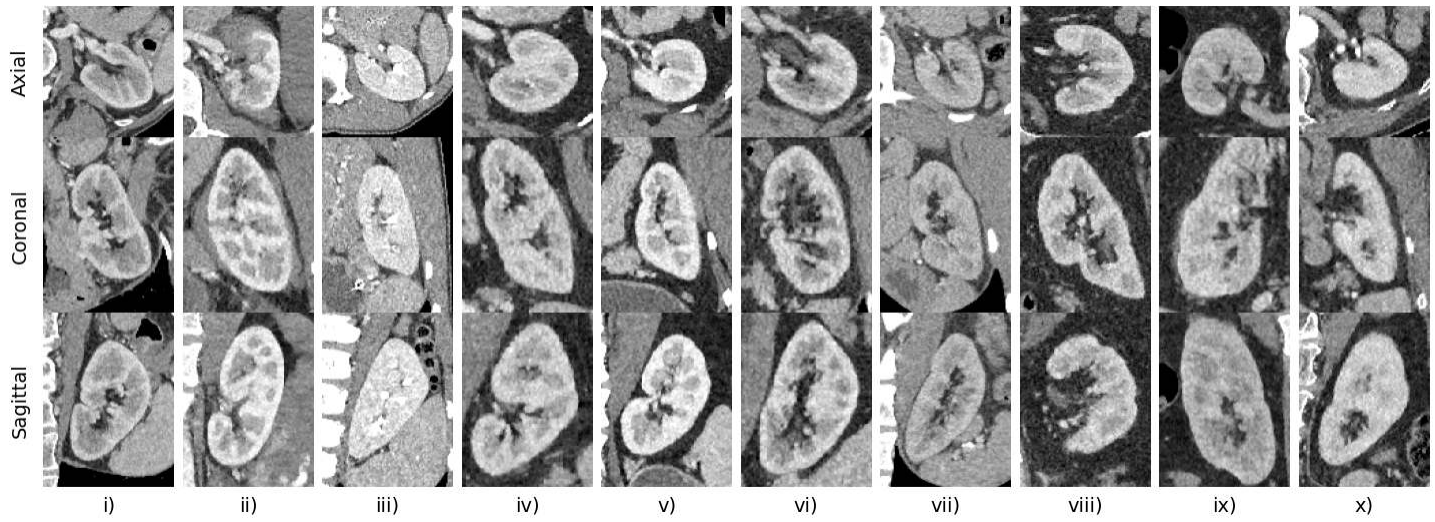
Parameter	**Value**
Diffusion steps	1000
Diffusion size	[32, 24, 24]
Diffusion channels	8
Dimension multipliers	[2, 4, 8, 16]
Optimizer	Adam
Learning rate	0.0001
EMA Decay	0.995
Loss	L1
Batch size	40
Gradient accumulation	Every 2 epochs
Maximum iterations	250,000

Table A.3: Classifier Hyperparameters

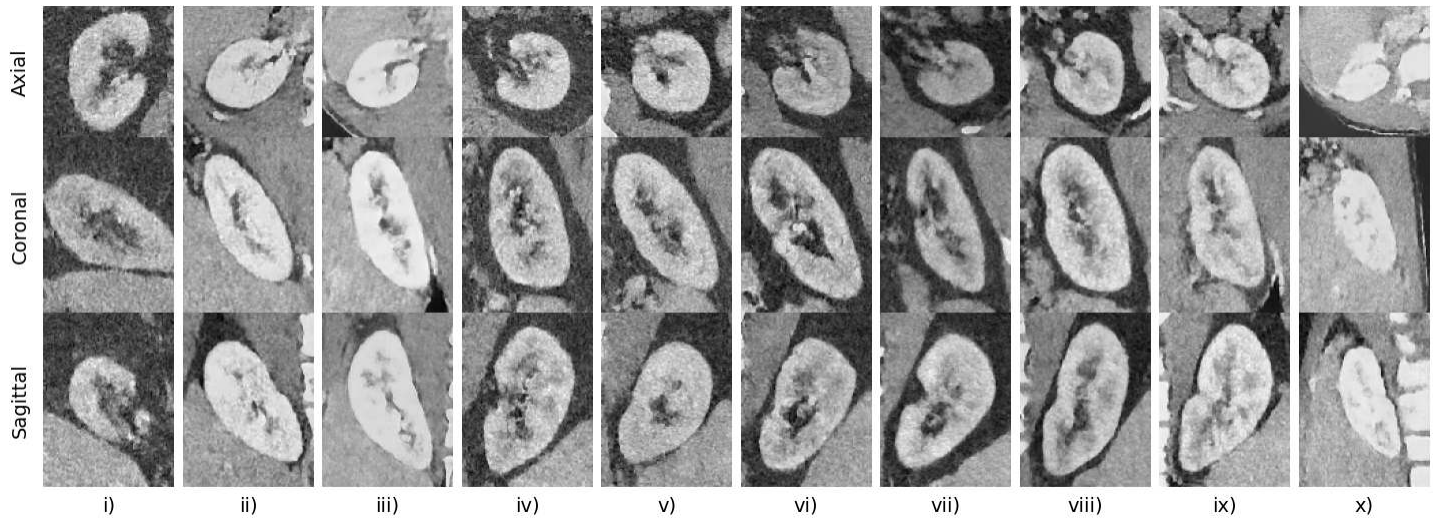
Parameter	**Value**
Backbone	
Base architecture	Same as Table A.2
Dropout	0.2
Frozen layers	2
Classification Head	
Layer sizes	[384, 192]
Dropout	[0.6, 0.36]
Optimizer	AdamW
Weight Decay	0.001
Learning rate	0.0001
Scheduler	CosineAnnealingWarmRestarts
Loss	Cross Entropy
Batch size	40
Gradient accumulation	Every 2 epochs

Appendix B

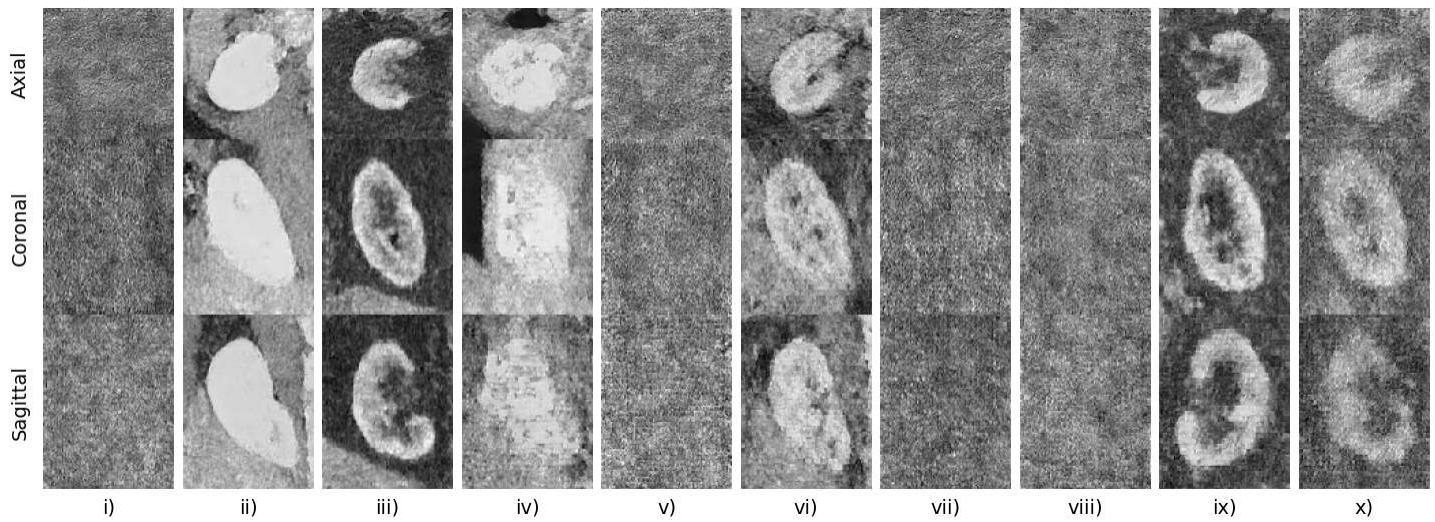
Generation Examples



(a) Real kidney CT patches



(b) DDPM generated kidney CT patches



(c) DDIM generated kidney CT patches

Figure B.1: **Generation examples of our diffusion model compared to real kidney CT patches.** a) shows real kidney CT patches, b) shows kidney CT patches generated by our DDPM model and c) shows kidney CT patches generated by our DDIM model.