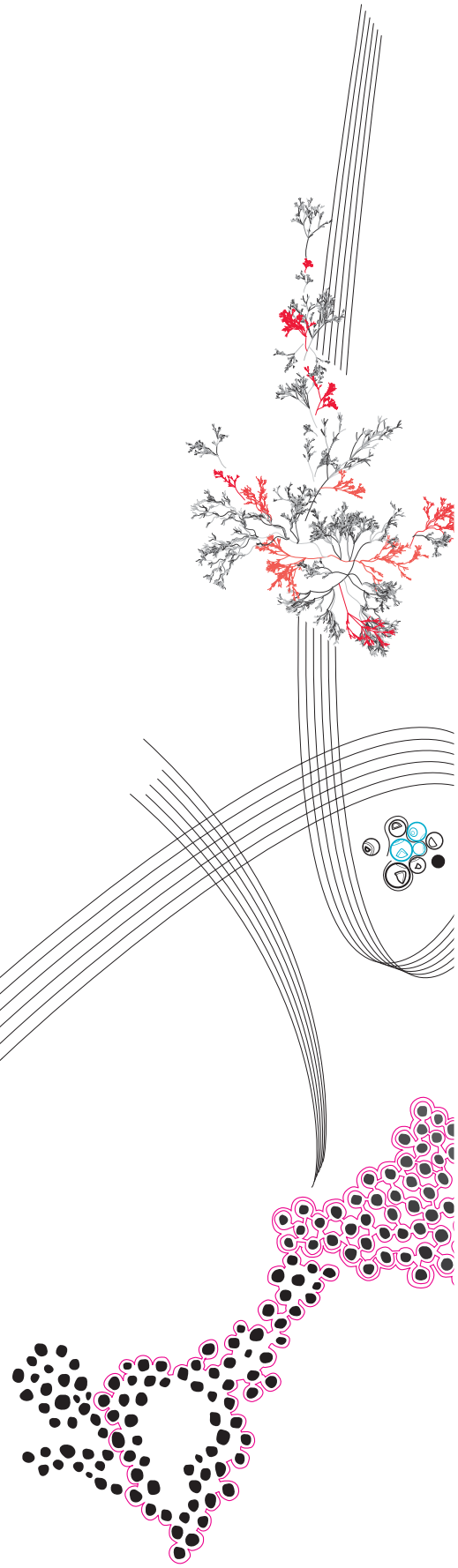

MSc Applied Mathematics
Final Project

Conditional Denoising Diffusion Probabilistic Models for Metal Artifact Reduction in Computed Tomography

Jorn Quattrocchi

Chair: prof. dr. Christoph Brune
Supervisor: dr. Jelmer M. Wolterink
External member: dr. Felix L. Schwenninger

External supervisor: ir. Jochen A. C. van Osch
External supervisor: dr. Mark Selles

June, 2025

Mathematics of Imaging & AI (MIA),
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Metal artifacts in computed tomography (CT) significantly degrade image quality, potentially introducing challenges in diagnostic evaluations. This study investigates the application of advanced diffusion-based deep learning models for metal artifact reduction (MAR) in CT images. Conditional Denoising Diffusion Probabilistic Models (CDDPM), Brownian Bridge Diffusion Models (BBDM), and their superconditional variant (SBBDM) were implemented in this research, trained on a large simulated dataset of CT slices with synthetic metal artifacts.

The comparative analysis includes an ablation study with the UNet backbone of the diffusion models, the clinically verified DL-MAR model, and the commercialized O-MAR method. Experiments on 30 clinical CT scans with unilateral hip prostheses demonstrate that diffusion models achieve superior artifact reduction compared to O-MAR and comparable performance to UNet-based approaches.

While diffusion models show promise in MAR applications, their computational exhaustion currently outweighs their occasional performance gains over more efficient UNet-based methods. This research provides insights into the strengths and limitations of diffusion-based approaches for MAR. Future research will be necessary to explore the potential clinical value of diffusion based techniques in the field CT imaging.

Keywords: Metal Artifact Reduction, Computed Tomography, DDPM, BBDM, Clinical Evaluation

Chapter 1

Introduction

Computed Tomography (CT) has revolutionized medical imaging by enabling non-invasive cross-sectional visualization of anatomical structures [8]. However, the presence of metallic implants introduces severe streaking artifacts through physical phenomena including beam hardening, photon starvation, and scatter effects [2]. These artifacts degrade image quality by up to 30% in adjacent soft tissues, potentially obscuring critical diagnostic information and compromising radiotherapy planning accuracy [28].

Traditional metal artifact reduction (MAR) techniques employ sinogram interpolation strategies such as normalized MAR (NMAR) [20] and frequency-split approaches such as FSNMAR [21]. Commercialized implementations of sinogram interpolation techniques like Orthopedic MAR (O-MAR) demonstrate some artifact reduction , but they struggle with complex implant geometries and often introduce secondary artifacts. [28]

Deep learning approaches using UNet architectures [25] marked a shift in the MAR research field, achieving superior performance through learned artifact representations. However, their clinical deployment faces a critical challenge. The absence of paired clinical training data necessitates reliance on simulated metal artifacts [41]. The proposed UNet-based model DL-MAR, by M. Selles, was trained on simulated data and clinically verified as a strict improvement to commercialized methods like O-MAR [29].

Denoising Diffusion Probabilistic Models (DDPMs) [11] offer interesting advantages through their iterative denoising process and inherent uncertainty quantification. By learning the manifold of artifact-free CT images through progressive denoising and conditioning on the artifact affected image, Conditional DDPMs achieve state-of-the-art performance in image restoration tasks [26]. Brownian Bridge Diffusion Models (BBDMs) [15] present an alternative approach through direct input-output domain mapping, though their mathematical foundation remains problematic for exact image-to-image translations.

This research implements and evaluates Conditional DDPM (CDDPM), BBDM and superconditional BBDM (SBBDM) architectures using a simulated dataset of 113,462 CT slices with synthetic metal artifacts [28]. A comparative experimental analysis with the UNet backbone of the diffusion models as an ablation study (UNET), clinically verified UNet-based DL-MAR and commercialized O-MAR was conducted in this research on 30 clinical CT scans with a unilateral hip prosthesis. It was demonstrated that diffusion models achieve superior artifact reduction with respect to the commercialized method (O-MAR), and comparable state-of-the-art performance with respect to UNet based models like the ablation study (UNET) and DL-MAR. The clinical validation in this research introduces a contra-lateral consistency metric and a pixel-wise variance map to quantify uncertainty. However, the performance of the diffusion based methods presented in this research do not outweigh their computational expense.

Chapter 2

Problem Description

Computed tomography (CT) is a widely used technique to quickly obtain cross-sectional images of a patient. The technique is based on shooting narrow x-ray beams through the patient, which are detected opposite to the x-ray source. Sophisticated mathematical algorithms then calculate a two-dimensional image of the slice of the patient. Metal implants in the imaged area of the body can cause so-called metal artifacts in the reconstructed image, which is a generic term for several types of image-corrupting effects caused by the metal present. In this chapter, computed tomography and metal artifacts will be introduced to provide context on this topic for the rest of this report. In the last section of this chapter, the generalized problem of this research will be mathematically formulated.

2.1 Computed Tomography

An illustration of a CT scanner is depicted in Figure 2.1. Photons are accelerated from the x-ray source towards a target. The number of photons that reach the detectors are registered as a one-dimensional projection of the target. The gantry is rotated to obtain projection data from multiple angles. [7]

When a photon travels through a medium, there is a probability that the photon will be absorbed by the medium or scattered off the original trajectory. The attenuation of the photons is dependent on the energy of the photons and the medium it is traveling through. This probability of attenuation is a medium characteristic, described by the Beer-Lambert law [12]:

$$I = \int I_0(E) e^{-\int \mu(E,s) ds} dE, \tag{2.1}$$

with intensity I of the photon beam measured by the detector, initial intensity $I_0(E)$ of the photon beam depending on the energy level E of the photons, linear attenuation coefficient $\mu(E, s)$ describing the probability of the photon attenuating over a unit distance depending on energy E and position s, the integral $\int ds$ over the line the photon beam traveled and integral $\int dE$ over the energy range over which the photon source generates photons.

The different attenuation characteristics of materials provide a way to differentiate between materials within a scan. The intensity of a pixel in a scan is typically represented in Hounsfield Units (HU):

$$HU = 1000 \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}, \tag{2.2}$$

which is the linear attenuation μ of the scanned material scaled such that air is -1000 HU and water is 0 HU. In a CT scan of the body, human tissue values will usually lie between

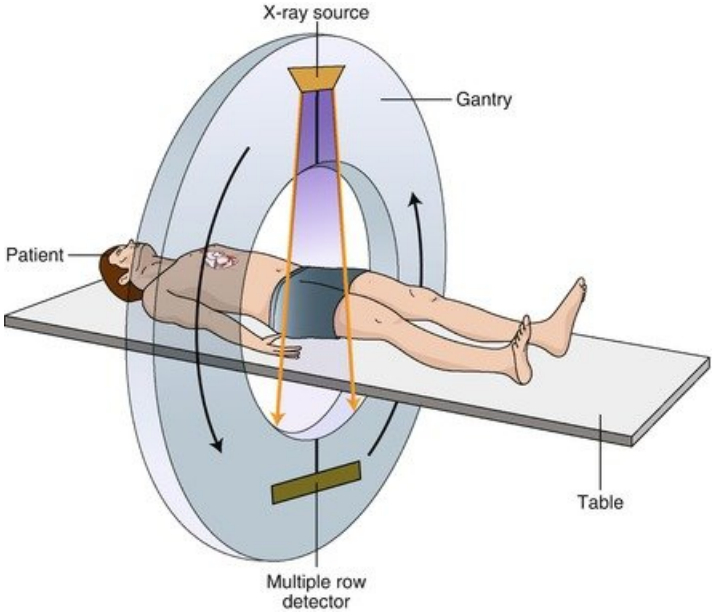

FIGURE 2.1: Figure demonstrating the main components of a CT machine, including the gantry, x-ray source, detector array, and the table for translating the patient. [7]

−700 HU (typical intensity of the lung) and 1800 HU (typical intensity of the cortical bone). [12]

The x-ray beams are shot at a fixed angle with the target. By repeating the projection process, conventionally over an equally spaced angular partition, projection data is obtained for multiple angles covering 360 degrees. This set of projection data is called a sinogram, from which the two dimensional image is reconstructed. An example of a sinogram of a simple oval body phantom can be found in Figure 2.2. In this sinogram one horizontal line is a one dimensional projection of a particular angle. The complete sinogram is a vertical stack of projections [12].

The mathematical foundation that relates the two-dimensional image domain to the sinogram domain is the Radon transform. The Radon transform $\mathcal{R}f$ is defined as integrating a function f on $\mathbb{R}^n$ over its hyperplanes [23]. In two dimensions, the Radon transform consists of parallel line integrals over all angles. Let an image be the result of a function $f(x_1, x_2)$ defined on $\mathbb{R}^2$, with Cartesian coordinates x_1 and x_2. Let a straight line L be defined by line coordinates ρ and θ, where ρ is the signed distance of the line to the origin and θ is the angle of the line with the x_1-axis. A set of parallel lines is easily constructed by fixing θ and varying ρ. Now Radon transform $\mathcal{R}$ of $f(x_1, x_2)$, denoted by $\hat{f}(\rho, \theta)$, is defined to be the line integral $\int d\tau$ for all straight lines defined by angle θ and displacement ρ:

$$\hat{f}(\rho, \theta) = (\mathcal{R}f)(\rho, \theta) = \int_{-\infty}^{\infty} f(\tau\cos(\theta) - \rho\sin(\theta), \tau\sin(\theta) + \rho\cos(\theta))d\tau, \tag{2.3}$$

with $0 \leq \theta < 2\pi, -\infty \leq \rho \leq \infty$.

A visual example of the formalization of the Radon transform can be found in Figure 2.3. Now it can be clearly seen that a computed tomography scan is a discretized physical way to measure the Radon transforms of particular image slices. In other words, a sinogram is a measurement of a Radon-transformed cross-section of a patient [23].

To obtain the two dimensional image from the measured sinogram, filtered backprojection (FBP) is typically used in clinical CT. This method provides a fast, simple and suffi-

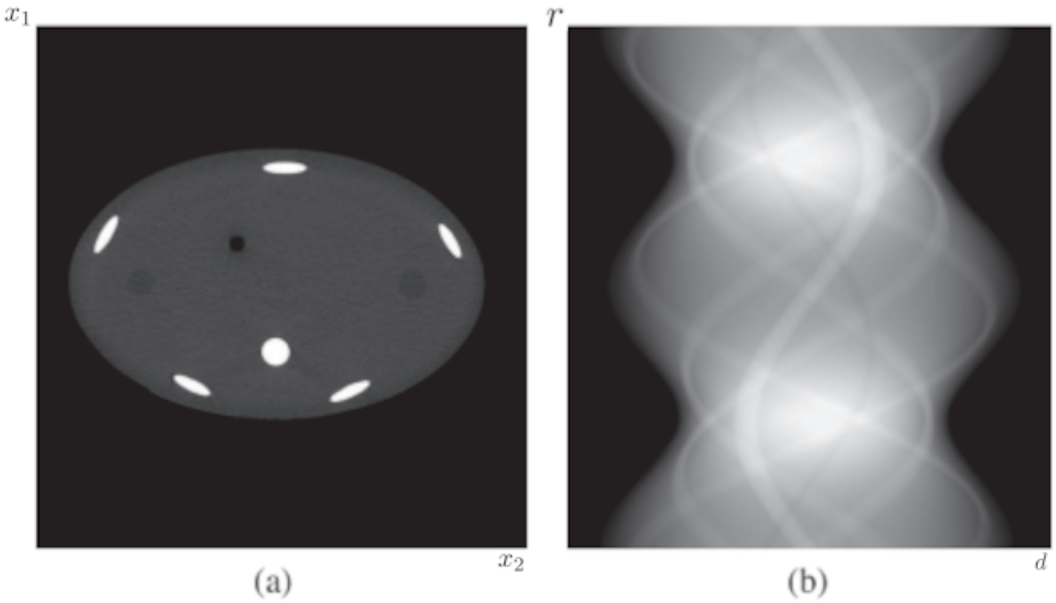

FIGURE 2.2: (a) An oval body phantom, represented in the two-dimensional Cartesian plane. (b) The corresponding sinogram, with the scan angles on the vertical axis and the detector bins on the horizontal axis. The horizontal width d represents the detector width of the scanner, i.e. the number of detector bins. Vertical length r is the full angular range of the scanner, typically and in this case set to $r = 2\pi$ for a full rotation. Resolution of (a) corresponds to the detector width in (b), $x_1 \times x_2 = d \times d$. [12]

ciently precise reconstruction, which are all highly desirable in clinical applications. The projection data is filtered to remove most of the noise and then back-projected from each projection angle to form the cross-sectional image. The basic idea of the back-projection is based on the inverse of the Radon transform and illustrated in Figure 2.4, One can see that each pixel in the reconstructed image is dependent on all x-ray paths that passed through the respective area (pixel) in the patient [12].

2.2 Metal Artifacts

When metal is present in the body, several types of artifacts can occur, corrupting the reconstructed image and reducing the diagnostic capacity. [2] An example of severe metal artifacts caused by bilateral hip prostheses can be found in Figure 2.5 (a).

Metal artifacts are typically dark and bright streaks originating from the implant. Due to the different types and sizes of metal present in the body, the severity of the artifacts can be very different. The artifacts are caused by a wide variety of phenomena. Poisson noise, beam-hardening and scatter effects are caused by the metal itself. The metal edges partially entering slices causes undersampling and motion artifacts [2]. Examples of the different artifacts are shown in Figure 2.5.

When inspecting the x-ray beam on a photon level, One can see that the beam actually consists of individually independent photons. Therefore, the photons arrive independently at the detectors following a Poisson process. Poisson noise is the statistical error caused by the independent arrivals of photons, thus a property of the signal itself. Statistical errors are larger when the sample size is smaller, so the Poisson noise is larger in detector bins

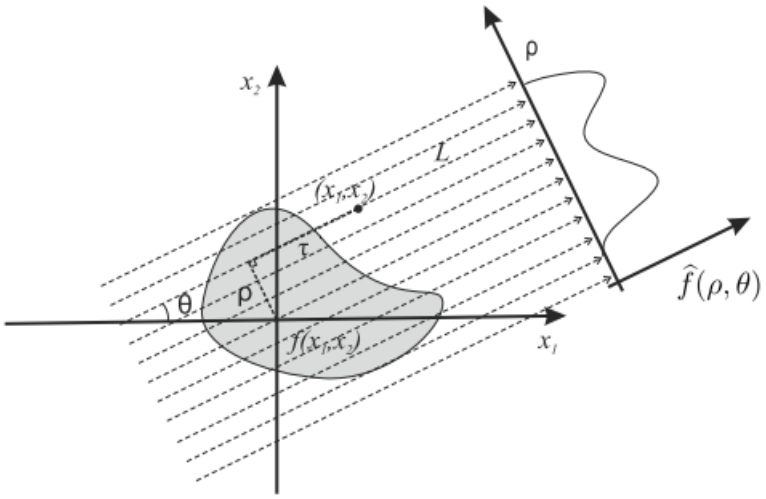

FIGURE 2.3: A two-dimensional object $f(x_1, x_2)$ and its projections $\hat{f}(\rho, \theta)$. Cartesian (x_1, x_2) and line (ρ, θ) coordinates are indicated. Continuously traveling over a line L via τ. [23]

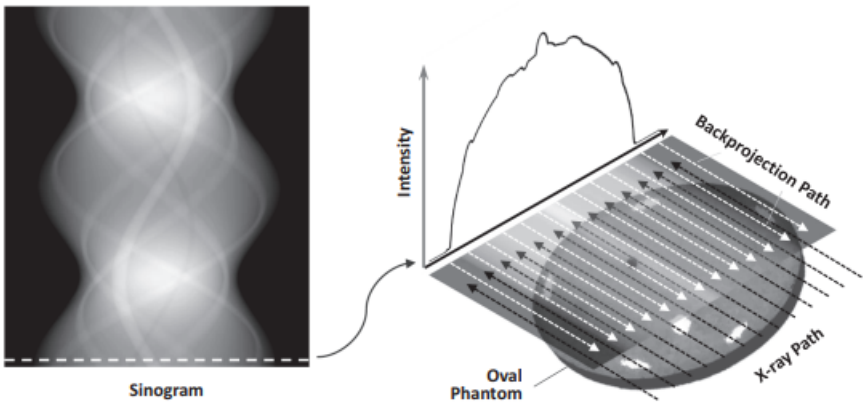

FIGURE 2.4: Illustration of the backprojection concept. Intensities of the measured sinogram represent the line integrals of the attenuation coefficients of the object along x-ray paths shown by the black dashed arrows. A measured sinogram intensity is backprojected along the exact x-ray path that produced the measurement. [12]

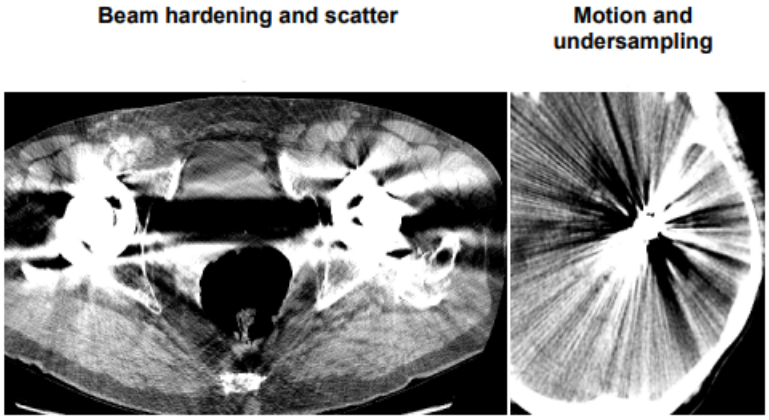

FIGURE 2.5: (a) Dark streak hip prosteses caused by beam-hardening and scatter. (b) Sharp thin altering streaks surrounding an aneurysm coil caused by motion and undersampling artifacts. [2]

with fewer photon counts. Poisson noise causes random thin dark and bright streaks. A high Poission noise may obscure soft tissue boundaries, while high contrast objects (metal, bone) are still visible [2][12].

Beam-hardening and scatter are two different physical phenomenon, both causing dark and bright streaks after reconstruction. When a high attenuating medium (high atomic number), like metal, is present in the body, low energy photons are attenuated more easily than high energy photons. This will lead to an x-ray beam primarily consisting of higher energy photons 'hardening' the beam, causing the beam-hardening effect. Scatter, or more precisely Compton scatter, is the effect caused by higher energy photons that attenuate by changing direction instead of absorption. Changing direction will sometimes mean that the photon ends up in a different detector than expected, this will lead to a large error when that detector would otherwise have very few photons. Both beam-hardening and scatter result in more photons being detected than expected, which causes the dark streaks. The bright streaks are a byproduct of the high pass filter in the FBP algorithm exaggerating differences between adjacent detectors [2][8][12].

When the desired area of a patient is moving while scanning, motion artifacts can cause blurring and long range streaks. The long range streaks are situated between high contrast edges and the position of the x-ray source at the time of movement, therefore metal present in the body can increase corruption by motion artifacts [2][10].

Undersampling is a problem for any discretized measurement of the continuous world. Ideally in computed tomography, the slice thickness and detector bin width approach zero to approximate the Radon transform. However, since this is clearly not reachable, undersampling artifacts can occur after image reconstruction. For example, the partial volume effect is a direct consequence of undersampling. The slice volume which is measured as if it was a plane has room for multiple tissues to partially enter the same pixel. To illustrate, when a high contrast volume partially enters the measured slice, the detector bins measure a mixed attenuation, blurring the image. This will cause more prominent artifacts with a higher contrast between tissues. The partial volume effect is often visible near small metals, like dental fillings, surgical clips and needles [2][21][13].

2.3 Problem Formulation

For any patient, an image of a single two dimensional slice $s \in \mathbb{R}^2$ is desired. For a finite set of angles $\Theta \subset \mathbb{R}$ and displacements $P \subset \mathbb{R}$, CT measurements of the Radon transformed slice $\mathcal{R}s$ are stored in sinogram $\tilde{f}(\rho, \theta)$ including measurement error $\epsilon \in \mathbb{R}$ for all $\theta \in \Theta, \rho \in P$:

$$\tilde{f}(\rho, \theta) = (\mathcal{R}s)(\rho, \theta) + \epsilon, \quad \text{with } \theta \in \Theta, \rho \in P, \epsilon \in \mathbb{R}. \tag{2.4}$$

The discretized measurements will result in a discretized representation $\tilde{s}(x_1, x_2)$ of slice s, with (x_1, x_2) from finite discrete set $X^2 \subset \mathbb{R}^2$. Observing from the Radon transform, the highest resolution without interpolation would be when the length of X is equal to the length of P, i.e. $|X| = |P|$.

Obtaining discretized approximation $\tilde{s}(x_1, x_2)$ of slice s from $\tilde{f}(\rho, \theta)$, can be formulated as a variational inverse problem:

$$\min_{\tilde{s} \in \mathbb{R}^2} D(\tilde{s}, \tilde{f}) + \alpha R(\tilde{s}), \tag{2.5}$$

with data fidelity term D and regularization term R weighted by scalar $\alpha \in \mathbb{R}$. The forward operator is the Radon transform $\mathcal{R}$, so the data fidelity term D should be a distance measure between $\mathcal{R}\tilde{s}$ and $\tilde{f}$. Because noise is typically present in the data, One should take a regularization term in consideration. Data fidelity term D and regularization term αR are not further specified, since the scope of this research is metal artifact reduction on top of existing reconstruction algorithms.

An algorithm that finds optimal solutions for a variant of the variational inverse problem (2.5), is from now on called the reconstruction algorithm Q. Taking as input the measured sinogram $\tilde{f}$ and giving a discretized optimized reconstruction $s^* \in \mathbb{R}^2$ of s as output:

$$Q(\tilde{f}) = s^* \tag{2.6}$$

For typical reconstruction algorithms, like the FBP algorithm, s^* could still be corrupted by metal artifacts $a \in \mathbb{R}^2$. The true clinical discretized representation of the desired slice $s(x_1, x_2) \in \mathbb{R}^2$ is without metal artifacts. Therefore, reconstructed image s^* could be written as a superposition of the true representation of the patient slice s and the metal artifacts a still present in s^*:

$$s^* = s + a \tag{2.7}$$

The metal artifact reduction problem can now be formulated. Find metal artifact disentanglement model M_Q for reconstruction algorithm Q, such that for any given image s^* reconstructed by Q:

$$M_Q(s^*) = s. \tag{2.8}$$

The actually used reconstruction algorithm is not specified further in the formulation, therefore the model M_Q suffices for any Q. It should be mentioned that the present metal artifacts in s^* are very dependent on the reconstruction algorithm, so a dedicated model M_Q should be considered per reconstruction algorithm Q.

Chapter 3

Theory and Related Work

Since CT was introduced in the 1970s, improving its image quality has been a popular research topic. The physical inconvenience of metal objects significantly attenuating or completely blocking x-rays remains a major problem in the reconstruction of CT images, which has received a lot of attention in the field of medical image reconstruction [8].

In the first section of this chapter the different approaches of conventional methods will be elaborated on. The remaining sections explain advances in machine learning and their application in metal artifact reduction. Within the domain of machine learning, several important image-to-image (I2I) techniques and generative models will be discussed. The last section will focus on state-of-the-art machine learning models and the scope of this research: diffusion models.

3.1 Conventional Metal Artifact Reduction

Obviously, One could choose for a very invasive method by removing the metal implants prior to the scan. The removal of dental fillings prior to the CT scan of the patient has been studied by Gray et al. [10]. Not surprisingly, this completely stopped metal artifacts from occurring, but is clearly not a realistic approach when the implants are hip prostheses, for example. Invasive surgery is in itself a problem with additional complications and non-metal alternatives to prostheses are often inadequate. Conventional MAR methods focus on correction techniques later in the image reconstruction process [8][29].

The first noninvasive moment metal artifacts can be reduced is during the scan. Adjusting the parameters to minimize metal artifacts has been investigated. For example, the photon energy in the x-ray could be increased to reduce beam-hardening and noise. Unfortunately, an increase in photon energy means an increase in radiation dose for the patient. Usually, adjusting parameters to increase image quality means an increase in radiation or it gives rise to other image corrupting effects. Overall, in standard CT this approach is not sufficient for many severe metal artifact cases. However, it should be mentioned that new x-ray tomography techniques, like dual energy computed tomography (DECT) [31] and spectral photon counting computed tomography (SPCCT) [33], show promising improvements towards image quality and reducing metal artifacts as well. Few facilities use these techniques in practice yet, so an improvement of MAR in standard CT is still desired [8].

Post-scan MAR techniques are based on improving the reconstruction technique and/or correcting the raw projection data. These techniques are non invasive and by far the most researched. The first projection based MAR technique was proposed by Kalender et al. in 1987 [13]. Their technique identified the metal trace in the sinogram domain

and linear interpolation was used to replace the data in the metal trace. This technique and other early techniques based on interpolation had notable disadvantages. Anatomical areas suffered from blurring and due to a lack of smoothness in the sinogram, interpolating caused new streaking artifacts [13].

Meyer et al. proposed normalized metal artifact reduction (NMAR) in 2010, which was a significant improvement to prior models [20]. In NMAR, the interpolation is computed in a normalized sinogram. First, a prior image is created by segmenting the image into basic tissues like air, soft tissue and bone. Then the sinogram was transformed into its normalized represenation, dividing the sinogram by the projection data of the prior image. As a result of the normalization, anatomical variations are minimized, while the metal trace in the sinogram is masked. When interpolation is performed in this normalized domain, it is less likely to create abrupt transitions or misrepresent underlying anatomical structures. Finally, the sinogram is denormalized to reintroduce the correct anatomical variations. Meyer et al. showed that this reduced blurring and new streaking artifacts substantially [20].

In 2012, the same authors improved their own model with the proposition of frequency-split normalized metal artifact reduction (FSNMAR) [21]. FSNMAR combines the uncorrected image with the corrected image from NMAR through filtering. They observed that artifacts caused by beam-hardening and scatter often have relatively low frequencies, so high-pass filtering the uncorrected image would extract edge information about anatomical structures. FSNMAR then combined the high-pass filtered uncorrected image with a more reliable low-pass filtered NMAR corrected image. Compared to NMAR, FSNMAR showed less blurring and improved depiction of anatomical structures near the metal [21].

Leading companies commercialized their own version of interpolation-based MAR, all inspired by NMAR or FSNMAR. Orthopedic MAR (O-MAR) by Philips, iterative MAR by Siemens Healthineers (iMAR), Single Energy MAR (SEMAR) by Canon and Smart-MAR (MARS) by GE Healthcare. The commercial algorithms have been extensively reviewed. Strong metal artifact reduction is achieved by these models. However, in the corrected image some artifacts may still be present corrupting small anatomical structures. Furthermore, additional secondary artifacts may be introduced [29].

3.2 Machine Learning

Machine learning has revolutionized the field of computer vision, enabling systems to automatically learn and make predictions from visual data. In the imaging field, machine learning models play a crucial role in tasks such as image classification, generation and object detection. Additionally, many problems in computer vision can be formulated as an image-to-image (I2I) translation task. Image segmentation, enhancement and denoising are exemplary tasks for which machine learning was successfully applied [26][41].

Machine learning typically consists of a network model M_θ that performs a desired approximation or generative task, with learnable parameters θ. A loss function L penalizes the model iteratively during training for incorrect predictions, updating θ accordingly by back-propagating the loss over the network.

Machine learning models can be categorized by their learning method. In **unsupervised learning**, the model is trained on an unlabeled dataset. This means that image models must identify the meaningful features without any explicit guidance. Unsupervised image-to-image translation advanced significantly with CycleGAN [43], which extended generative adversarial networks (GANs) [9] by introducing bidirectional cyclic consistency.

In **supervised learning**, the algorithm is trained on a labeled dataset. Each data

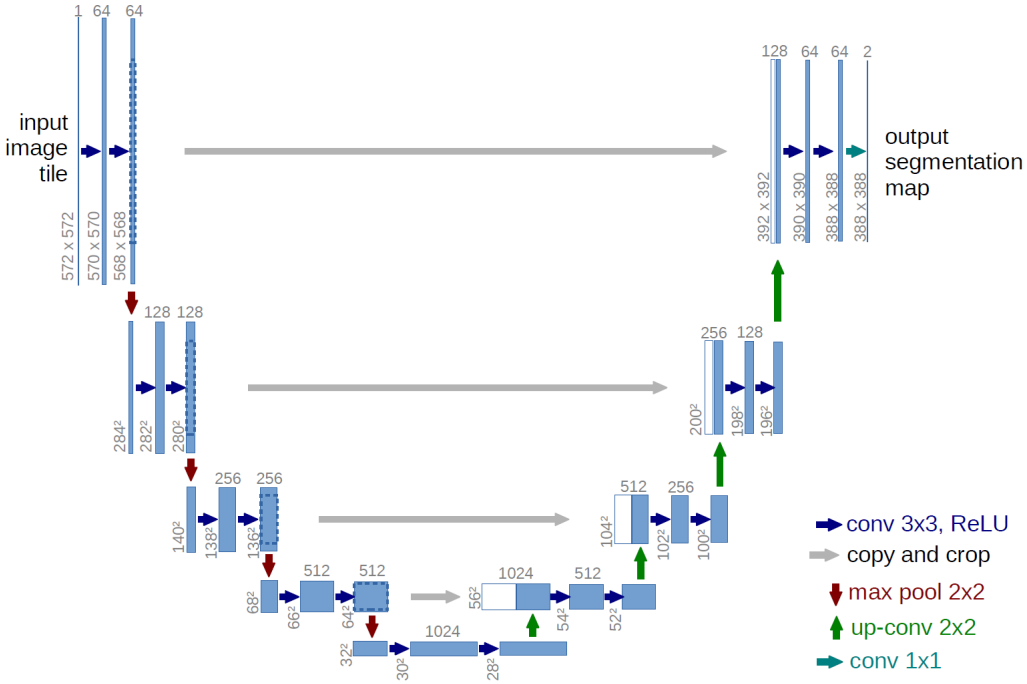

FIGURE 3.1: The structure of a U-Net, as proposed by Ronneberger et al. The left and downward contracting encoder path with convolutional and max pooling steps learns to represent the input image in features, stored in an increasing number of feature channels. The right and upward expanding decoder path with the addition of up-convolutions steps learns to visualize the learned features by decoding the features into higher resolution pixel representations using the horizontal skip connections for spatial information. [25]

point, like an image, in the training set is paired with a corresponding label. This label can be anything the model should predict, like a class or another image. Image models learn to map the input images to their respective labels during the training process by identifying meaningful features. Once trained, the model should be able to predict the labels of unseen images. For image-to-image translation tasks, deep convolutional architecture types are a popular supervised approach. The deep convolutional U-Net architecture was a major breakthrough in I2I translation. The U-Net was first proposed by Ronneberger et al. for biomedical image segmentation. [25]

3.2.1 UNET

A U-Net consists of a contracting path with encoder blocks and an expansive path with decoder blocks. The model gets its name from the symmetric design, since the encoder and decoder blocks can be represented as a U-shaped architecture. In the encoder blocks, images are convoluted to extract features and downsampled via pooling operations. The number of feature channels increase in each encoding block. In the decoder blocks, the feature channels are upsampled and convoluted, reducing the number of feature channels. Each decoder block gets spatial information from the corresponding higher resolution encoder block on the other side of the U-shape via skip connections. The diagram of the originally proposed U-Net architecture can be found in Figure 3.1. [25]

The original U-Net model could achieve good results with a limited amount of data.

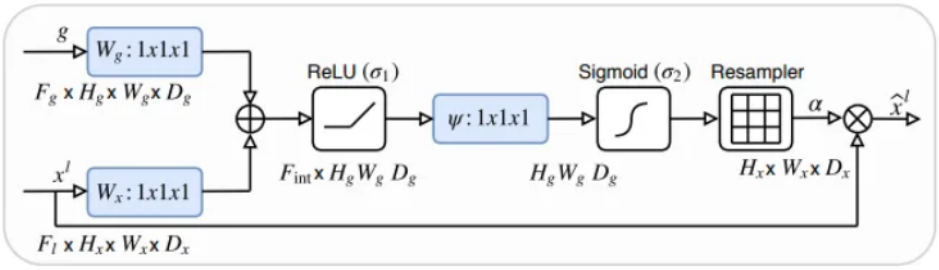

FIGURE 3.2: Schematic of an additive attention gate (AG). Input features (x^t) are scaled with attention coefficients (α) computed in AG. Spatial regions are selected by analyzing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale. Grid resampling of attention coefficients is done using trilinear interpolation. [22]

However, some applications required more depth in the U-Net framework to achieve the desired performance. Simply adding more layers, deepening the U-Net, resulted in a drastic increase in training time or a degradation of accuracy. Most subsequent U-Net architectures are based on additional residual components and attention mechanisms to the skip connections, to be able to upscale the model efficiently and improve accuracy. Since the publication of Ronneberger et al. many improvements have been made to the original U-Net architecture. ResUNet [5], RU-Net [32], Attention UNet [22], AResU-Net [40] and Residual-Attention UNet++ [17] are some of the wide variety of publications proposing an improved U-Net architecture based on residual and attention learning.

Residual Attention

Deep neural networks are prone to vanishing gradients, which residual blocks help mitigate. Residual components follow the form:

$$H(x) = F(x) + x \tag{3.1}$$

where $F(x)$ represents learned transformations. This design enables direct gradient flow through identity mappings. [5]

The attention mechanism, as proposed in Attention U-Net [22], introduces learnable gating to focus on relevant image regions. An attention gate (Figure 3.2) computes coefficients α that weight encoder features x based on both local features and global contextual information g. Vector g is the output from the next lower resolutional layer, so the features in g are better represented. The vector x is the higher dimensional output from the encoder on the other side of the U-Net and goes via the skip connection into the attention gate. After convolutions to bring the vectors to the same dimensions, the vectors are summed element-wise. The summed vectors go through more activation functions and convolutions to end up with attention coefficients α ranging between 0 and 1. A higher attention coefficient indicates more relevancy. The attention coefficients α are multiplied element wise with the encoder features x. The attention weighted spatial information will go in the next decoder together with the output g of the lower dimensional decoder and will be handled like in a normal U-Net.

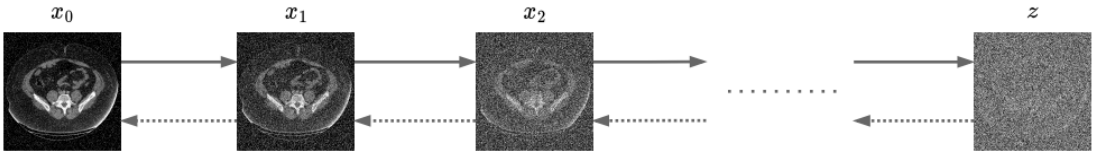

FIGURE 3.3: Forward ($x_0 \rightarrow z$) and backwards ($x_0 \leftarrow z$) process of diffusion models.

Timestep Embedding

Diffusion models, introduced in the next section, require timestep conditioning of the denoising backbone to guide the iterative denoising process across timesteps with time dependent noise levels.

The timestep embedding process uses sinusoidal positional encoding, which maps discrete timesteps t to continuous vector representations through frequency-modulated sine and cosine functions. For a timestep t and embedding dimension $i \in \{0, ..., d-1\}$:

$$\gamma(t)^{(i)} = \begin{cases} \sin\left(\frac{t}{10000^{2k/d}}\right) & \text{if } i = 2k \\ \cos\left(\frac{t}{10000^{2k/d}}\right) & \text{if } i = 2k+1 \end{cases} \tag{3.2}$$

where d is the embedding dimension and $k \in \{0, ..., d/2-1\}$. This produces a d-dimensional vector $\gamma(t) \in \mathbb{R}^d$. The vector representation $\gamma(t)$ preserves relative time relationships, because neighbouring timesteps have smoothly varying embeddings [11].

3.3 Diffusion Models

Although being a recent addition to the generative field, diffusion models have proven to be a valuable approach across various applications [11] [1] [35]. The diffusion models are already widely adopted by society through applications like the pioneering DallE by OpenAI.

This section offers an exploration of the mathematical theory of diffusion models, covering the fundamental diffusion model Denoising Diffusion Probabilistic Model (DDPM) [11] and an alternative approach for conditional diffusion modeling Brownian Bridge Diffusion Model (BBDM) [15]. The DDPM is motivated from a variational perspective, then BBDM is formulated analogously.

3.3.1 Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) were proposed by Ho et al. in 2020 [11]. The DDPMs consist of a forward and reverse process. The forward process is a deterministic Markov chain to construct latent variables for each data sample by gradually adding Gaussian noise at each time step. The reverse process approximates the inverse to obtain a sample from the data distribution, starting from pure Gaussian noise. The diffusion process is visualized in Figure 3.3.

Let $q(x_0)$ be the distribution of the noise-free data, latent variables $x_1, ..., x_T$ are produced for sample $x_0 \sim q(x_0)$ by adding Gaussian noise at time t as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \forall_{t\in\{1,...,T\}}, \tag{3.3}$$

with T the number of time steps in the diffusion process, $\beta_1, ..., \beta_T \in [0, 1)$ defining the variance schedule for each timestep and I is the identity matrix. The normal distribution is represented by $\mathcal{N}(x; \mu, \sigma)$ with mean μ and covariance σ. Since this forward process is a deterministic Markov chain, the transition to any x_t can be formulated directly conditioned on x_0. Let $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, then the one step formulations for the latent distributions are

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1 - \overline{\alpha}_t)I), \forall_{t\in\{1,...,T\}}. \tag{3.4}$$

Now for any input x_0 One can sample a noise component $\epsilon \sim \mathcal{N}(0, I)$ to obtain any latent variable x_t as follows:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + (1 - \overline{\alpha}_t)\epsilon. \tag{3.5}$$

With sufficiently large diffusion length T and according variance schedule $\beta_1, ..., \beta_T$, the latent distribution $q(x_T)$ at the end of the diffusion process should be approximately equal to the normal distribution:

$$q(x_T) \approx \mathcal{N}(x_T; 0, I). \tag{3.6}$$

Therefore the reverse process is modeled by starting at Gaussian noise and can be formulated accordingly. Starting at $p(x_T) = \mathcal{N}(x_T; 0, I)$ the approximated reverse process is formulated as follows:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{3.7}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3.8}$$

The goal for this model is that $p_\theta(x_0)$ approximates the true data distribution $q(x_0)$. Which means that the parameters θ should be optimized such that it maximizes the likelihood that the generated data samples from $p_\theta(x_0)$ belong to $q(x_0)$. However, using the negative log-likelihood $-log(p_\theta(x_0))$ as a loss function is not desirable. Calculating the log-likelihood for high dimensional data with continuous values (like images) over T time steps is intractable. To simplify the objective function the authors of DDPM take several steps, starting with the variational lowerbound on the negative log likelihood:

$$\mathbb{E}[-\log(p_\theta(x_0)] \leq \mathbb{E}_q\Big[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\Big] = \mathbb{E}_q\Big[-\log p(x_T) - \sum_{t\geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\Big] =: L \tag{3.9}$$

This loss function L is then rewritten in time step components $L_{vlb} = L_0 + L_1 + ... + L_T$, using the Kullback-Leibler divergence (D_{KL}) as the statistical distance between distributions. The derivation can be found in appendix A from the publication by Ho et al. [11], the results follow here:

$$L_{vlb} = L_0 + L_1 + ... + L_T, \tag{3.10}$$

with

$$L_0 := \mathbb{E}_q\Big[- \log p_\theta(x_0|x_1)\Big] \tag{3.11}$$

$$L_{t-1} := \mathbb{E}_q\Big[D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\Big], \forall_{t\in\{2,...,T\}} \tag{3.12}$$

$$L_T := \mathbb{E}_q\Big[D_{KL}(q(x_T|x_0)||p(x_T))\Big]. \tag{3.13}$$

By experimental evidence of the authors the L_0 component was discarded. The posterior q could have learnable parameters, however by fixing the forward process variances β_t to sufficiently small constants relative to the data domain, the authors ensured an approximately equal functional form in the forward and reverse process while maintaining $q(x_T)$ to be sufficiently close to the normal distribution. This causes L_T to be a constant during training, so L_T could also be discarded. Now L_{vlb} only consists of the intermediate time step components L_{t-1} with $t \in \{2, ..., T\}$. These loss functions directly compare $p_\theta(x_{t-1}|x_t)$ against forward process posterior distributions $q(x_{t-1}|x_t, x_0)$ via the Kullback-Leibler divergence. This means that the models task has been reduced to estimating the parameters of $q(x_{t-1}|x_t, x_0)$, such that these Kullback-Leibler divergences in equations (3.12) are minimized. The posterior distribution can be written down in the same form as equation (3.8):

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{3.14}$$

with

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t \tag{3.15}$$

$$\tilde{\beta}_t := \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t. \tag{3.16}$$

Since the authors decided to fix the variances β_t to constants, equation (3.8) can be reformulated by setting $\Sigma_\theta(x_t, t) = \sigma_t^2 I$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I). \tag{3.17}$$

With equations (3.14) and (3.17) with constant variances, minimizing the Kullback-Leibler divergences of the loss functions in (3.12) come down to simply minimizing the distance between the means of two Gaussian distributions. This can be done as follows:

$$L_{t-1} = \mathbb{E}_q\Big[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2\Big] + C, \forall_{t\in\{2,...,T\}}, \tag{3.18}$$

where C is a constant independent of θ. There are now several approaches to parameterize μ_θ, the first one is to predict x_0 directly and find μ_θ through equation (3.15). The second approach is to predict the forward process posterior mean $\tilde{\mu}_t(x_t, x_0)$ completely, which would be the most intuitive approach. However, the authors of DDPM find a more elegant third approach. Equation (3.4) can be reparameterized as $x_t(x_0, \epsilon) = \sqrt{\overline{\alpha}_t}x_0 + (1-\overline{\alpha}_t)\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, which means with equation (3.15) the loss function could be written as:

$$L_{t-1} - C = \mathbb{E}_{x_0,\epsilon}\Big[\frac{1}{2\sigma_t^2}\Big|\Big|\frac{1}{\sqrt{\alpha_t}}\Big(x_t(x_0, \epsilon) - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon\Big) - \mu_\theta(x_t(x_0, t), t)\Big|\Big|^2\Big]. \tag{3.19}$$

In this equation it can be clearly seen that $\mu_\theta(x_t(x_0,t),t)$ should predict something depending on x_t. Since x_t is an input to the model, $\mu_\theta(x_t(x_0,t),t)$ could be parameterized equivalently:

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t(x_0,\epsilon) - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t,t)\Big). \tag{3.20}$$

Here $\epsilon_\theta(x_t,t)$ is a function approximator with the task to predict ϵ from x_t, so the model is reduced to finding the noise component of x_t. Resulting in the following loss function:

$$L_{t-1} - C = \mathbb{E}_{x_0,\epsilon}\Big[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\overline{\alpha}_t)}\Big|\Big|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\Big|\Big|^2\Big]. \tag{3.21}$$

A proof on differentiability of equation (3.21) and a more extended deduction can be found in the publication by Ho et al. [11]. However, the main contribution of this paper is a simplified version of equation (3.21):

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\Big[\Big|\Big|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\Big|\Big|^2\Big]. \tag{3.22}$$

The authors of DDPM showed that this simple alternative to equation (3.21) was beneficial for sampling quality. Now the model could just be trained on the simple objective to minimize the mean squared distance between the noise component $\epsilon \sim \mathcal{N}(0,I)$ drawn in the forward process and the models prediction of that noise $\epsilon_\theta(x_t,t)$. The complete training algorithm can be found in Algorithm 1.

Algorithm 1 Training DDPM

1: **repeat**
2: Draw $x_0 \sim q(x_0)$.
In this step, a sample from the to be learned data distribution is selected.
3: Draw $t \sim \text{Uniform}(\{1,...,T\})$.
In this step, a timestep t is randomly selected.
4: Draw $\epsilon \sim \mathcal{N}(0,I)$.
In this step, a Gaussian noise component is drawn.
5: Calculate $\overline{\alpha}_t = \prod_{s=0}^{t} 1 - \beta_s$
In this step, the cumulative product $\overline{\alpha}_t$ is calculated depending on the predefined variance schedule β_t and timestep t.
6: Take gradient descent step on:
$\nabla_\theta||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2$
In this step, the image x_0 from the to be learned distribution $q(x_0)$ and noise component ϵ are scaled and summed to calculate x_t. Then the noise prediction from the model $\epsilon_\theta(x_t,t)$ is compared with the actual noise component ϵ in the loss function (here the l_2 norm). Finally the gradient descent step is taken.
7: **until converged**

The reverse diffusion process or sampling procedure always starts with a random Gaussian sample $x_T \sim \mathcal{N}(0,I)$. Then x_{t-1} can be iteratively sampled from the distribution $p_\theta(x_{t-1}|x_t)$, which is the same as computing $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t,t)\Big) + \sigma_t z$ with $z \sim \mathcal{N}(0,I)$. The complete sampling algorithm can be found in Algorithm 2.

Algorithm 2 Sampling DDPM

1: $x_T \sim \mathcal{N}(0, I)$
In the sampling procedure x_T is drawn as Gaussian noise.
2: **for** $t = T, ..., 1$ **do**
Traveling backwards over the diffusion bridge, the image x_t is refined for each step.
3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
A Gaussian noise component z is drawn for each $t > 1$, else $z = 0$.
4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$
Then x_{t-1} is calculated, using trained predictor ϵ_θ to predict the noise component in x_t.
5: **end for**
6: **return** x_0

3.3.2 Conditional DDPM

DDPM is a powerful generative framework, outperforming GANs in sample quality and training stability. The DDPM model can learn the distribution of a training dataset well enough to generate convincing samples during inference. However, image-to-image (I2I) translation tasks require the possibility to condition on an input image. The impressive generalizing qualities of DDPM inspired many publications to adapt the DDPM model for I2I translation problems.

The model UNIT-DDPM [27] by Sasaki et al. proposed a dual-domain connected Markov chain model based on DDPM. The two diffusion chains in each domain learned standard denoising, like in DDPM, and translating partially denoised images between domains. In inference, noise was added like in the forward diffusion of DDPM until a certain release time was reached. Then the input image plus noise was translated to the Markov chain in the other domain to be denoised like in the reversed diffusion of DDPM.

In Palette [26], proposed by Saharia et al. the input image was added as a condition to the denoising network in each time step during training and inference. The model outperforms GANs on colorization, inpainting, uncropping and JPEG restoration. Examples of the results of Palette for each task can be found in Figure 3.4.

A publication by [34] Wolleb et al. adapts the approach of Palette for medical image segmentation, where the input image is an additional condition for the denoising network at each time step.

MedSegDiff [35], proposed by Wu et al. and SegDiff [1] proposed by Amit et al. also propose conditioned diffusion models for segmentation tasks. MedSegDiff and SegDiff propose a similar approach, where the conditional image is first encoded into features before adding the conditional features in the bottleneck of the denoising network at each time step.

Brownian Bridge Diffusion Models (BBDMs)

In comparison to most existing diffusion models, Li et al. propose an I2I translation framework that directly builds the mapping between the input and output domains via a Brownian bridge [15]. An illustration of the Brownian bridge process can be found in Figure 3.5. In this section, formulations will be as consistent as possible with the previous section on DDPMs.

A Brownian bridge is a stochastic model in which the start and end states are given, the probability distribution during the diffusion process is conditioned on those start and

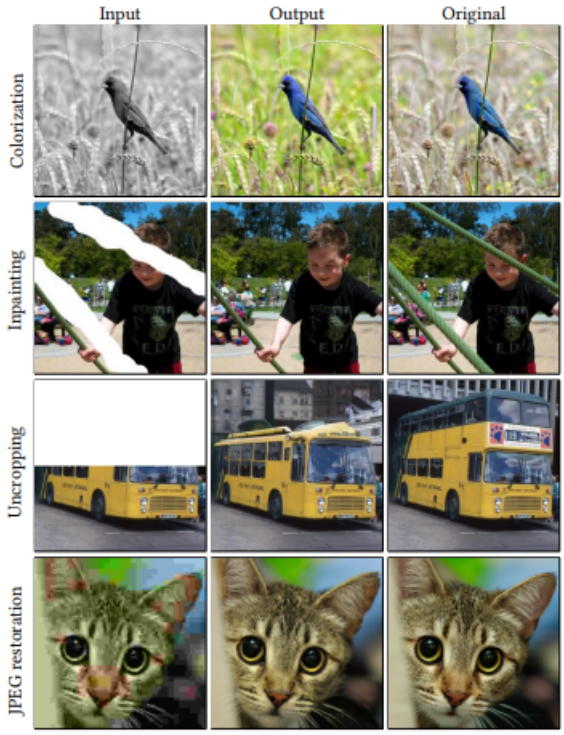

FIGURE 3.4: Image-to-image diffusion model Palette generates high-fidelity samples for colorization, inpainting, uncropping and JPEG restoration. [26]

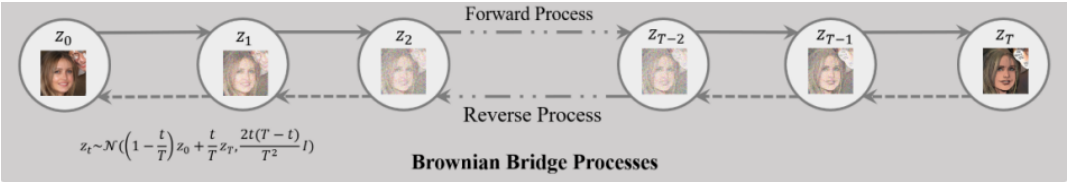

FIGURE 3.5: Forward and reverse Brownian Bridge process of the Brownian Bridge Diffusion Model (BBDM). [15]

end states. The state distribution of intermediate states on the Brownian bridge, starting from $x_0 \sim q(x_0)$ at $t = 0$ and ending at x_T at $t = T$, can be formulated as:

$$p(x_t|x_0, x_T) = \mathcal{N}((1 - \frac{t}{T})x_0 + \frac{t}{T}x_T, \frac{2t(T-t)}{T^2}I). \tag{3.23}$$

The formulation shows that the diffusion process is fixed at both ends with zero variance. The process in between is a Brownian bridge, with most variance in the middle of the bridge $t = T/2$.

Let (x, y) be a paired data point from the training dataset. Images x and y are from the image domains A and B, respectively. The corresponding forward diffusion process of BBDM can now be defined as:

$$q_{BB}(x_t|x_0, y) = \mathcal{N}(x_t; (1 - m_t)x_0 + m_t y, \delta_t I), \forall_{t\in\{1,\dots,T\}}, \tag{3.24}$$

with $x_0 = x$, $x_T = y$, $m_t = t/T$ and T is the number of timesteps. The variance schedule δ_t is designed by the authors as follows:

$$\delta_t = 2s(m_t - m_t^2), \tag{3.25}$$

with scaling factor s. This variance schedule is not following the definition from equation (3.23), because this will induce a maximal variance at $t = \frac{T}{2}$ or $\delta_{T/2} = \frac{T}{4}$. This would become an extremely large variance for large T, so the variance schedule δ_t is designed to have a maximal variance of:

$$\delta_{T/2} = \frac{s}{2}. \tag{3.26}$$

The maximal variance in BBDM is therefore defined by s. This scaling factor s is set to 1 by default, the influence of s will be discussed in the next chapters.

Now for any input (x_0, y) and timestep $t \in \{1, ..., T\}$ One can sample a noise component $\epsilon_t \sim \mathcal{N}(0, I)$ to obtain any latent variable x_t as follows:

$$x_t = (1 - m_t)x_0 + m_t y + \sqrt{\delta_t}\epsilon_t. \tag{3.27}$$

For the reverse process, BBDM sets $x_T = y$. Input y is available during training in the data pair (x, y) and during inference as the conditioning image for the generative process. The goal is now to predict x_{t-1} given x_t, the approximated reverse process is formulated as follows:

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \tilde{\delta}_t I) \tag{3.28}$$

with predicted mean $\mu_\theta(x_t, t)$ and $\tilde{\delta}_t$ the variance of the noise. Like in DDPM, $\mu_\theta(x_t, t)$ will be learned via maximum likelihood criterion. One can observe that only for $t = T$, the condition y is actually a condition for the distribution p_θ.

Analogously to DDPM, the model is trained by optimizing the variational lower bound L_{vlb}. This is constructed and rewritten in time step components using the Kullback-Leibler divergence:

$$L_{vlb} = L_0 + L_1 + L_2 + ... + L_T, \tag{3.29}$$

with

$$L_0 = \mathbb{E}_{q_{BB}}[-\log p_\theta(x_0|x_1, y)], \tag{3.30}$$

$$L_{t-1} = \mathbb{E}_{q_{BB}}[D_{KL}(q_{BB}(x_{t-1}|x_t, x_0, y)||p_\theta(x_{t-1}|x_t, y))], \forall_{t\in 2,...,T}, \tag{3.31}$$

$$L_T = \mathbb{E}_{q_{BB}}[D_{KL}(q_{BB}(x_T|x_0, y)||p(x_T|y))]. \tag{3.32}$$

The L_0 term is discarded analogously to DDPM. Since in BBDM x_T is equal to y, the L_T term is constant and can be discarded. The remaining terms are handled by obtaining the distribution formula for $q_{BB}(x_{t-1}|x_t, x_0, y)$:

$$q_{BB}(x_{t-1}|x_t, x_0, y) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0, y), \tilde{\delta}_t I), \tag{3.33}$$

with

$$\tilde{\mu}_t(x_t, x_0, y) = \frac{\delta_{t-1}}{\delta_t}\frac{1 - m_t}{1 - m_{t-1}}x_t + (1 - m_{t-1}\frac{\delta_{t|t-1}}{\delta_t})x_0 + (m_{t-1} - m_t\frac{1 - m_t}{1 - m_{t-1}}\frac{\delta_{t-1}}{\delta_t})y \tag{3.34}$$

and variance term

$$\tilde{\delta}_t = \frac{\delta_{t|t-1}\delta_{t-1}}{\delta_t}. \tag{3.35}$$

An elaborate deduction can be found in Appendix A of BBDM [15]. During inference x_0 is unknown, so similarly to DDPM $\tilde{\mu}$ is reparametrized, by substituting equation (3.27) into equation (3.34). By reformulating equation (3.27) we get

$$x_0 = \frac{x_t - m_t y - \sqrt{\delta_t}\epsilon_t}{1 - m_t}. \tag{3.36}$$

Substituting this expression of x_0 into equation (3.34) gives us

$$\tilde{\mu}_t(x_t, x_0, y) = \tilde{\mu}_t(x_t, y) = \frac{\delta_{t-1}}{\delta_t}\frac{1 - m_t}{1 - m_{t-1}}x_t \tag{3.37}$$

$$+(1 - m_{t-1}\frac{\delta_{t|t-1}}{\delta_t})(\frac{x_t - m_t y - \sqrt{\delta_t}\epsilon_t}{1 - m_t}) + (m_{t-1} - m_t\frac{1 - m_t}{1 - m_{t-1}}\frac{\delta_{t-1}}{\delta_t})y.$$

The authors of BBDM state that equation (3.34) is equal to:

$$\tilde{\mu}_t(x_t, y) = c_{xt}x_t + c_{yt}y + c_{\epsilon t}(m_t(y - x_0) + \sqrt{\delta_t}\epsilon), \tag{3.38}$$

with the introduction of the following constants.

$$c_{xt} = \frac{\delta_{t-1}}{\delta_t}\frac{1 - m_t}{1 - m_{t-1}}, \tag{3.39}$$

$$c_{yt} = (m_{t-1} - m_t\frac{1 - m_t}{1 - m_{t-1}}\frac{\delta_{t-1}}{\delta_t}) \tag{3.40}$$

and

$$c_{\epsilon t} = (1 - m_{t-1}\frac{\delta_{t|t-1}}{\delta_t}). \tag{3.41}$$

To verify this, it should be possible to get from equation (3.34) to equation (3.38). First the constants c_{xt}, c_{yt} and $c_{\epsilon t}$ can be substituted directly into equation (3.34):

$$\tilde{\mu}_t(x_t, y) = c_{xt}x_t + c_{\epsilon t}x_0 + c_{yt}y. \tag{3.42}$$

Now, One could see easily that the new expression of $\tilde{\mu}_t$ in equation (3.38) is only true if $x_0 = m_t(y - x_0) + \sqrt{\delta_t}\epsilon$. This is clearly problematic.

For example, when $t = T$, recalling $m_t = \frac{t}{T}$ and $\delta_t = 2s(m_t - m_t^2)$, the equation (3.38) is true if and only if $x_0 = \frac{1}{2}y$.

In this research, the gap in the BBDM paper is ignored in the application of the model. This will also be brievely mentioned in the discussion of this research, chapter 6.

The authors of BBDM simplify the loss function with respect to their representation of $\tilde{\mu}_t(x_t, y)$. This simplified loss function for BBDM is as follows:

$$L_{simple} = \mathbb{E}_{x_0,y,\epsilon}[c_{\epsilon t}||m_t(y - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, t)||^2], \tag{3.43}$$

where the constant $c_{\epsilon t}$ is discarded in the training algorithm, so:

$$L_{simple} = \mathbb{E}_{x_0,y,\epsilon}[||m_t(y - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, t)||^2]. \tag{3.44}$$

Now predictor $\epsilon_\theta(x_t, t)$ is the network to be trained. The BBDM paper calls this the noise predictor, but that is not completely true. The predictor does more than that. Recalling equation (3.27), L_{simple} can be reformulated in a more readable way:

$$L_{simple} = \mathbb{E}_{x_0,y,\epsilon}[||x_t - x_0 - \epsilon_\theta(x_t, t)||^2]. \tag{3.45}$$

Now it is obvious that the predictor needs to estimate the difference between x_t and x_0. This difference consists of the noise component and the mean difference component:

$$x_t - x_0 = \sqrt{\delta_t}\epsilon + \tilde{\mu}_t(x_t, x_0, y) - x_0. \tag{3.46}$$

The proposed training and sampling algorithm can be found in Algorithm 3 and 4, respectively.

Algorithm 3 Training BBDM

1: **repeat**
2: $x_0 \sim q(x_0), y \sim q(y)$
In this step, a training pair x_0 and y is randomly selected from data distributions $q(x_0), q(y)$.
3: $t \sim \text{Uniform}(\{1, ..., T\})$
In this step, a timestep t is randomly selected.
4: $\epsilon \sim \mathcal{N}(0, I)$
In this step, a Gaussian noise component is drawn.
5: Take gradient descent step on:
$\nabla_\theta ||m_t(y - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, t)||^2$
In this step, the prediction of predictor ϵ_θ is compared with the noise component plus the mean difference component via the l_2 norm. Then the gradient decent step is taken.
6: **until converged**

Algorithm 4 Sampling BBDM

1: $x_T = y \sim q(y)$
In the sampling procedure x_T is selected from the known distribution $q(y)$.
2: **for** $t = T, ..., 1$ **do**
Traveling backwards over the Brownian bridge, the image x_t is moving in a noisy path towards the desired distribution $q(x_0)$.
3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
A Gaussian noise component z is drawn for each $t > 1$, else $z = 0$.
4: $x_{t-1} = c_{xt}x_t + c_{yt}y + c_{\epsilon t}\epsilon_\theta(x_t, t) + \sqrt{\tilde{\delta}_t}z$
Then x_{t-1} is calculated.
5: **end for**
6: **return** x_0

Chapter 4

Methods

To find the added value of diffusion frameworks on metal artifact reduction, two models are considered. Both models are trained on a simulated paired dataset. For validation of the results a clinical dataset is obtained from the Isala hospital in Zwolle. The first model is a conditional adaptation of DDPM, as done by Wolleb et al. [34] and Saharia et al. [26]. The second model is based on the idea of the BBDM model by Li et al. [15], leveraging on the concept of directly mapping the input and target domain by a Brownian bridge. In this chapter, the datasets and models explored in this research will be discussed in detail to ensure reproducibility.

4.1 Dataset

4.1.1 Simulated Data

To obtain a large amount of paired training data for metal artifact reduction, simulating metal artifacts is required. Metal-free images were obtained from the open source Deep Lesion dataset [37]. The metal artifact simulation method was first proposed by Zhang et al. [41] and adapted by Selles et al. [28]. An overview of the simulation process can be found in Figure 4.1.

A subset of 2279 patients from the Deep Lesion dataset were included, with a total of 113,462 CT-images of 512x512 pixels. Manual segmentation has been done to acquire 35 different metal masks, consisting of smaller implants like surgical metal clips and larger implants like hip prostheses.

Different kinds of metal implants are bound to different anatomical areas. To automatically find the corresponding anatomical area in a CT-image, a residual neural network was trained by Selles et al. with an accuracy of 99.2% on the validation dataset.

For a certain metal free CT-image x, the residual neural network finds an implant from the 35 metal masks to fit with the anatomy in the image x. To determine the location of the metal mask, a bone probability map is obtained to ensure the metal implant is placed near or in bone structures. Now a mask x_m is constructed for the desired implant in the desired location.

Then beam-hardening and Poisson noise artifacts are simulated as proposed by Zhang et al. [41]. By using the principles of CT, projection data can be simulated by computing line integrals simulating x-ray projections. To be able to compute the line integrals, in each pixel the material has to be known to determine attenuation coefficients. Therefore, the CT-image x is divided in a water image x_w and bone image x_b by a soft threshold-based method [41]. Projection data of x_w, x_b and x_m are obtained by simulation, using

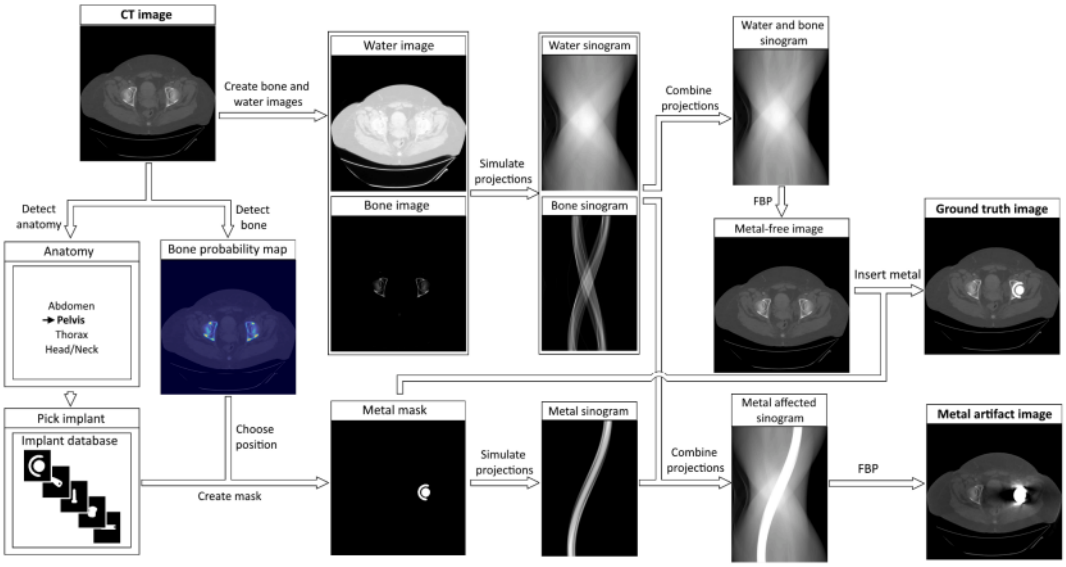

FIGURE 4.1: A flowchart of the simulation of metal artifacts in an artifact-free CT image slice, proposed by Selles [29]. This algorithm allows the artificial construction of a paired dataset to train a MAR model.

the attenuation coefficient of iron in place of the metal mask. A metal affected sinogram is obtained by first combining the water sinogram and bone sinogram, then all nonzero values in the metal sinogram are substituted into the metal affected sinogram. Finally, FBP is used to get the metal artifact image x_a. Similarly, the ground truth image x_g is obtained by combining the water and bone sinogram. The metal-free image is obtained again by FBP, then the metal mask is used to replace the values in the location of the implant by CT values of iron.

For each of the 113,462 Deep Lesion CT-images, the simulation process is applied to obtain 113,462 data pairs with a ground truth image and a metal artifact image. The obtained simulated paired dataset is divided by patient into a train, validation and test partition pursuing a $80 : 10 : 10$ split. The exact division of the dataset can be found in Table 4.1.

TABLE 4.1: Partition of simulated dataset.

Partition	Patients	Image Pairs	%
Train	1852	91547	80.7
Validation	210	11477	10.1
Test	217	10438	9.2

4.1.2 Clinical Data

For evaluation purposes, a clinical dataset is obtained from the Isala hospital in Zwolle. Thirty patients with a unilateral hip prosthesis that were scanned between August 2022 and April 2023 are included retrospectively in this dataset. The patients were scanned on a Philips Spectral CT 7500 system.

4.2 Conditional Denoising Diffusion Probabilistic Model

By conditioning on the artifact affected image, as done by Wolleb et al. [34] and Saharia et al. [26], the DDPM framework is adapted for the I2I problem of metal artifact reduction.

The training and sampling algorithms for conditional DDPM (CDDPM) change slightly, by adding the conditional image as an input to the denoising network in each timestep. The algorithms used in this research for training and sampling are found in Algorithm 5 and 6 respectively.

Hyperparameters

The CDDPM model consists of the following hyper parameters:

- Number of timesteps T
- Noise schedule $\beta_t, \quad \forall t \in \{0, ..., T\}$
- Denoising network ϵ_θ

The denoising network will be later discussed in its own section.

Algorithm 5 Training CDDPM

1: **repeat**
2: paired data $x_0 \sim q(x_0), y \sim q(y)$
3: $t \sim \text{Uniform}(\{1, ..., T\})$
4: $\epsilon \sim \mathcal{N}(0, I)$
5: Take gradient descent step on:
$\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, y, t)||^2$
6: **until converged**

Algorithm 6 Sampling CDDPM

1: input $y \sim q(y)$
2: $x_T \sim \mathcal{N}(0, I)$
3: **for** $t = T, ..., 1$ **do**
4: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
5: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, y, t)\Big) + \sigma_t z$
6: **end for**
7: **return** x_0

4.3 Superconditional Brownian Bridge Diffusion Model

The BBDM training and sampling algorithms from the previous chapter, Algorithm 3 and 4 respectively, are implemented identically to BBDM and applied to metal artifact reduction.

Similarly to CDDPM, the BBDM denoising network can be conditioned on the artifact affected image in each timestep. Since BBDM is already conditioned on an input image, the additionally conditioned BBDM model is therefore called Superconditional BBDM (SBBDM). This model comes with slightly altered training and sampling algorithms, which can be found in Algorithm 7 and 8 respectively.

Hyperparameters

The (S)BBDM model consists of the following hyper parameters:

- Number of timesteps T.
- Maximal variance parameter s, defining the noise schedule δ_t.
- Denoising network ϵ_θ.

The denoising network will be later discussed in its own section.

Algorithm 7 Training SBBDM

1: **repeat**
2: $x_0 \sim q(x_0), y \sim q(y)$
3: $t \sim \text{Uniform}(\{1, ..., T\})$
4: $\epsilon \sim \mathcal{N}(0, I)$
5: Take gradient descent step on:
$\nabla_\theta ||m_t(y - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, y, t)||^2$
6: **until converged**

Algorithm 8 Sampling SBBDM

1: $x_T = y \sim q(y)$
2: **for** $t = T, ..., 1$ **do**
3: $z \sim \mathcal{N}(0, I)$ if $t > 1$, else $z = 0$
4: $x_{t-1} = c_{xt}x_t + c_{yt}y + c_{\epsilon t}\epsilon_\theta(x_t, y, t) + \sqrt{\tilde{\delta}_t}z$
5: **end for**
6: **return** x_0

4.4 Denoising Network Architecture

The denoising network ϵ_θ is yet to be defined. To be able to compare the models CDDPM and (S)BBDM, the denoising network will be defined once and used consistently. However, the denoising network is trained separately for each model and experiment.

A residual attention U-Net with timestep embedding is used, this follows the theory in section 3.2.1. To be consistent with BBDM and other related work, the widely used implementation of OpenAI [24] is adapted. When not mentioned in this paper, the hyperparameters are fixed and consistent with the adaptation of BBDM. [1]

4.5 Drawing Multiple Samples (n)

The diffusion based models discussed in this chapter, CDDPM and (S)BBDM, have inherent stochasticity due to the random noise component. During inference, the models predict an artifact-free image by iterative denoising, conditioned on a corresponding artifact-affected image. Due to the random noise component that is used, multiple samples can be drawn to inspect the variability in the models' predictions, like in the proposed method of Wolleb et al. [34]

[1] https://github.com/xuekt98/BBDM

For example, in MAR applications, repeated sampling could reveal if artifacts are systematically over- or undercorrected. Additionally, repeated sampling could show if the regions near the implants, where the information is most scarce, are inconsistently reconstructed.

In the experimental phase of this research, an extension of the sampling algorithm is considered by drawing n independent sample predictions $\{x_0^{(i)}\}_{i=1}^n$. The pixel-wise average $\overline{x}_0 = \frac{1}{n}\sum_{i=1}^n x_0^{(i)}$ can then be computed for a more stable prediction.

The *Manifold Hypothesis*, states that it can be presumed that a data distribution from the real world (like CT images) in a high-dimensional space (like the corresponding image space) lies on a low-dimensional manifold. When the desired data distribution is learned by a model, it can be assumed that any prediction x_0 from the model is on the lower dimensional manifold that covers the learned data distribution. However, without any knowledge about the convexity of the manifold, it cannot be assumed that $\overline{x}_0$ is on the manifold as well. Therefore, the added value of $\overline{x}_0$ will be discussed when evaluating the results of this research. [18]

4.6 Evaluation Criteria

For the evaluation of the models, a validation/test partition of the simulated dataset can be considered. In this dataset the ground truth is available, which makes an objective analysis possible. However, the data is synthetic and can be considered as an insufficient representation of real patient data. The clinical dataset can be used to showcase the value of the models in clinical applications. A ground truth is not available in this dataset. Therefore, evaluation criteria have to be defined for both datasets.

On the simulated dataset, the artifact-free prediction of the models can be compared to the ground truth with the Structural Similarity Index Measure (SSIM). The SSIM metric is closer to the human interpretation of similarity in images than standard metrics like the l_1 or l_2 norm. [42] The ultimate goal is to give a human doctor an artifact-free image to assess, so human interpretation is important. Therefore, SSIM is a suitable metric and will be used to evaluate the performance of the models on the simulated data.

On the clinical dataset, evaluating the artifact-free predictions of the models is not so straight forward due to the lack of ground truth data. Nevertheless, the following criteria will be used to evaluate the performance of the models on the clinical dataset:

1. Subjective analysis.
2. Contra-lateral image consistency.
3. Region of Interest (ROI) analysis.

Since the clinical dataset consists of only 30 patients, a subjective analysis is feasible. The goal is to find abnormal behaviour of the models and find a better understanding of possible shortcomings of the models. Considering bone or soft tissue hallucination, secondary artifacts and remaining artifacts.

The clinical dataset consists of patients with a unilateral hip prosthesis, so by the nature of metal artifacts the contra-lateral side of the image (without prosthesis) is less corrupted. To quantify the consistency of the anatomy in the artifact-free prediction by the models, the contra-lateral side of the input image can be compared to the same side of the prediction image with the SSIM metric. Compared to other verified methods, significantly different contra-lateral consistency can numerically verify that a model is hallucinating. However,

since the metal artifacts could reach the contra-lateral side, the optimal contra-lateral consistency is unknown. Therefore, this evaluation criterion is very limited in drawing conclusions.

The last evaluation criterion is based on the prior knowledge of the body. Regions of Interest (ROIs) can be defined to be in certain types of body tissue. One ROI in the bladder, one ROI in muscle tissue and one ROI in fatty tissue will be defined close to the prosthesis for each patient. Contrast-to-noise ratios between the bladder and fat, and between muscle and fat will be obtained to evaluate an improvement in readability of the predicted images. The contrast-to-noise ratios (CNRs) between the bladder and fat, as well as between muscle and fat, are determined by subtracting the average signal in the lowest signal tissue (fat) from the average signal in the other considered tissue (bladder or muscle), then dividing by the mean noise value of both tissues.

Chapter 5

Experiments and Results

For all experiments, a model m is trained for a maximal number of epochs E_m, so that convergence is achieved at epoch $e > 0$ with $e << E_m$. The maximal number of epochs E_m is picked by subjective experimental evidence. Explicitly,

$$E_m = 50 * \frac{1}{df_m},$$

where data fraction df_m is the fraction of the training dataset on which model m is trained. This is necessary to ensure the same maximal number of training iterations throughout all experiments.

The epoch e^*, where the model is performing best on the validation dataset, is considered optimal. This is done without the complete sampling procedure since this is a very computationally exhaustive process. Like in the training algorithm a random timestep t is uniformly drawn from $\{1, ..., T\}$, then the prediction of the estimator e_θ is tested against the ground truth images and included in the overall performance of the model on the validation dataset.

Due to the fact that diffusion models learn a distribution, the effect of drawing multiple samples is investigated as explained in section 4.5. To be able to evaluate the effect of repeated sampling, drawing 1, 5 and 10 independent sample predictions is considered for the diffusion models CDDPM and (S)BBDM. As promised in section 4.5, the average of these repeated sample predictions is computed as well and evaluated. Theoretically, the average will be more stable, but it should be verified as a feasible correction of an artifact-affected CT image.

5.1 Benchmark models

The presented experimental models in this research are benchmarked by comparing the predictions of experimental models with the predictions of the commercialized metal artifact reduction algorithm O-MAR (Philips) and clinically validated UNET based model DL-MAR (PhD. M. Selles) in the experiments on clinical data.

UNET

A second UNET model is trained and included in the experiments as an *ablation study*. The same UNET architecture is used for the denoising in the (S)BBDM and CDDPM models, but without the timestep embedding. This enables the investigation of the effect of the diffusion component in the experimental diffusion models. The UNET models are

trained on $df = 100, 10, 1$ percent of the synthetic training dataset. The models are named accordingly:

- UNET ($df = 100\%$)
- UNET ($df = 10\%$)
- UNET ($df = 1\%$)

5.2 CDDPM Experiments

For CDDPM models, the estimator ϵ_θ estimates the drawn noise component ϵ. The l_2-norm is conventionally used in diffusion models, so during inference this metric is also used to obtain the distance between the noise component and the estimated noise of the CDDPM models.

The CDDPM models in the experiments are conditioned on the artifact affected image. So we consider the Conditional DDPM (CDDPM) as a framework. These models are trained on $df = 100, 10, 1$ percent of the simulated training dataset to be able to assess the robustness of the models. The models are named accordingly, where df stands for data fraction:

- CDDPM ($df = 100\%$)
- CDDPM ($df = 10\%$)
- CDDPM ($df = 1\%$)

5.2.1 Synthetic Data Results

For computational reasons, only 50 data points from the test dataset are considered to be able to perform the whole 1000 step sampling procedure.

The overall results can be found in Table 5.1, supported by multiple figures in Appendix A. Figure A.1 to A.3, show a boxplot for each CDDPM model, showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

The UNET ($df = 100\%$) model is performing slightly better than the CDDPM models on the synthetic dataset. The UNET models are included in this diffusion research as an *ablation study*, since the same UNET architecture is used as a denoising backbone in the diffusion based models. On the synthetic validation data, there seems to be no added value by the diffusion component of CDDPM models.

By increasing the number of sample predictions drawn during inference, the performance of the average prediction increases for CDDPM.

Training on a Fraction of the Dataset (df)

The stochasticity of the CDDPM models should enable a higher generalizing capacity compared to the UNET models. Furthermore, it could be that the training dataset is too densely populated. A strong generalizing capacity would then be redundant, since the model is trained on a sufficient coverage of the input domain. The hypothesis would be that the UNET models would not be able to outperform the robustness of the CDDPM models when the population of the training data is less dense. Therefore, experiments on the size of the training dataset were included ($df = 100\%, 10\%, 1\%$) for UNET and CDDPM.

TABLE 5.1: Performance (SSIM) of the CDDPM and UNET models, trained on 100, 10 and 1 percent of the train partition of the simulated dataset. The performance was measured on 50 data points from the test partition. Each diffusion based model's performance is measured on the average of 1, 5 and 10 drawn samples during inference. In the first column the trained models are listed, in the second column the number of drawn samples are shown with the corresponding performance in the final column.

Model	**Samples**	**Performance (SSIM)**
CDDPM ($df = 100\%$)	1	0.9965
	5	0.9972
	10	0.9973
CDDPM ($df = 10\%$)	1	0.9965
	5	0.9972
	10	0.9972
CDDPM ($df = 1\%$)	1	0.9956
	5	0.9960
	10	0.9960
UNET ($df = 100\%$)	-	**0.9983**
UNET ($df = 10\%$)	-	0.9975
UNET ($df = 1\%$)	-	0.9956

The UNET models behave as expected, reducing the size of the dataset has a negative influence on the performance. The CDDPM models trained on 100 and 10 percent of the train partition of the synthetic dataset perform very similar. The CDDPM model trained on 1 percent of the training data shows that the CDDPM model slightly outperforms the UNET model trained on 1 percent.

The generalizing capacity of CDDPM actually seems to make the size of the dataset less important. This could be a motivation for diffusion based models in other image-to-image related research where training data is scarce. A clinical validation would strengthen this claim.

5.2.2 Clinical Data Results

The subjective evaluation is done for 5 of the 30 patients in the clinical dataset, by selecting a single slice containing a significant portion of the metal implant for each patient. Then the artifact-free predictions of each CDDPM model is compared with the predictions from the benchmark models O-MAR and DL-MAR.

The objective evaluation is done for all 30 patients in the clinical dataset, by selecting a single slice containing a significant portion of the metal implant for each patient. Then the artifact-free predictions of each CDDPM model and the predictions of the benchmark models are tested on the objective evaluation criteria.

Predictions

As mentioned, 10 sample predictions are obtained from the CDDPM models. The average of these 10 sample prediction is recalled to as the average prediction of the CDDPM models. All separate sample predictions of the diffusion based models can be found in Appendix B for a single image slice of 5 patients with a hip prosthesis.

In Figure 5.4, the (average) predictions of the CDDPM, UNET and benchmark models O-MAR and DL-MAR can be compared for the 5 clinical cases. In Figure 5.5 and 5.6, the (average) predictions for CDDPM and UNET are shown respectively. This double presentation of the results will help the reader navigate through the substantial number of experimental models.

Subjective Analysis

In the subjective performance analysis, the predictions of the models were analyzed on *hallucinations*, *secondary artifacts* and *remaining artifacts*. Any additional abnormal behavior will also be discussed in this section.

All CDDPM and UNET models show no clear sign of hallucinating bone structures or reshaping of soft tissue, which is a very good sign towards clinical applicability. The CDDPM models also show no hallucinations in the details of the average predictions. One could argue that some models may be hallucinating very close to the metal. However, due to the large disturbance by the metal artifacts close to the metal and the absence of a ground truth, it is not possible to confidently say how the bone should be restored underneath the heavy artifacts. Additionally, the shape of the metal is probably not correct in most cases, but this is of less importance with respect to the anatomical areas in the image.

All CDDPM and UNET models show no sign of secondary artifacts, which is a strict improvement from the commercialized model O-MAR.

All CDDPM and UNET models show some remaining artifacts, both the bigger shadow cast by the metal implant and the long thinner artifact streaks are occasionally present. The models CDDPM ($df = 100\%$) and UNET ($df = 100\%$) show most remaining artifacts in their predictions. The models CDDPM ($df = 1\%$), UNET ($df = 1\%$) show the least remaining artifacts and are outperforming both O-MAR and DL-MAR on this criterion.

The CDDPM and UNET model trained on 1% of the training data are convincingly outperforming their respective identical model trained on 100% of the training data, this contradicts the performance on the synthetic dataset.

Stochastic Analysis

For the CDDPM models, one could also calculate the standard deviation for each pixel, which could be seen as an uncertainty heat-map. In Figure 5.9, the average prediction, standard deviation and the average prediction with highlighted standard deviation is shown for 1 clinical case for each CDDPM model. The set of sample predictions for all 5 patients and all experimental diffusion models are included in Appendix B.

In the pixel wise standard deviation image for the CDDPM models, only artifacts and the negative shape of the metal can be found. Therefore, the models seem to be very capable of identifying the areas in the image which are affected by the metal artifacts. While the models are not so consistent in restoring these affected areas, drawing multiple samples enabled the quantification of their uncertainty. Identifying artifacts and quantifying uncertainty are both very important aspects towards clinical applications.

Unfortunately, consistently making the same error is still an error and this would not be visible in the pixel-wise uncertainty heat map. Therefore, the pixel wise uncertainty heat map should only be used to understand the models behavior and not directly to support clinical diagnostic decisions. This will be discussed in the limitations of this research.

Contra-lateral Consistency

For each of the 30 patients in the clinical dataset, the consistency between the image affected by artifacts and the artifact-free prediction on the contra-lateral side relative to the implant was evaluated (SSIM). The contra-lateral consistency of the average predictions from experimental models CDDPM ($df = 100\%$), CDDPM ($df = 10\%$), CDDPM ($df = 1\%$) and the predictions from the models UNET ($df = 100\%$), UNET ($df = 10\%$), UNET ($df = 1\%$) together with the predictions of the benchmark models O-MAR and DL-MAR can be found in Figure 5.12.

The O-MAR model is more consistent in the contralateral side than DL-MAR, while it was clinically verified that DL-MAR outperforms OMAR. This implies that the contralateral consistency is not essentially positively correlated with performance. The MAR models should be able to reduce minor artifacts in the contralateral side with respect to unilateral hip prostheses, so the optimal contralateral consistency depends on the severity of the artifacts. However, a very low contralateral consistency is still a sign of hallucinations. Additionally, this metric could help interpret the behavior of the experimental models.

The model CDDPM ($df = 100\%$) is even more consistent in the contralateral side than O-MAR, this suggests that in the predictions of this model most minor artifacts are still present in the contralateral side. While inspecting the models respective column of Figure 5.5, One could see that for all included patients a large portion of the artifacts still remain. This supports the claim that CDDPM ($df = 100\%$) undercorrects the contralateral side of patients with a unilateral hip prosthesis.

The models CDDPM ($df = 10\%$), CDDPM ($df = 1\%$), UNET ($df = 100\%$), UNET ($df = 10\%$) and UNET ($df = 1\%$) score very similar to DL-MAR on contralateral consistency. Looking at the respective columns in Figures 5.5 and 5.6, it can be seen that the correction of the minor artifacts in the contralateral side has improved with respect to CDDPM ($df = 100\%$).

Image Quality

For each of the 30 patients in the clinical dataset, three circular regions of interest (ROI) were drawn in the selected image slice by a clinically experienced PhD candidate. One was placed in the bladder at the medial side of the hip prosthesis, another was placed at the lateral side within the muscle area exhibiting the most pronounced artifacts, and a third was placed in the gluteal subcutaneous fat adjacent to the hip prosthesis.

The bladder-fat and muscle-fat CNRs are calculated for the average predictions from experimental models CDDPM ($df = 100\%$), CDDPM ($df = 10\%$), CDDPM ($df = 1\%$) and the predictions from the models UNET ($df = 100\%$), UNET ($df = 10\%$), UNET ($df = 1\%$) together with the predictions of the benchmark models O-MAR and DL-MAR. The results for these models can be compared in Figures 5.13 and 5.14. These Figures show boxplots for the bladder-fat and muscle-fat CNRs, respectively.

A higher contrast to noise ratio means that the different anatomical areas are clearly separable and less corrupted by noise. The box plots cover a wide range of values, which shows inconsistent performance for different patients. Therefore, model specific conclusions are limited on CNR.

The evaluation of the CNRs show that the models CDDPM and UNET trained on 100% of the training data are underperforming with respect to DL-MAR. All other CDDPM and UNET models score similar or even better than DL-MAR. This suggests that the predictions of these experimental models are at least as readable as the predictions of DL-

MAR. The best performing models on this evaluation metric are CDDPM ($df = 1\%$) and UNET ($df = 1\%$).

5.3 BBDM Experiments

For the BBDM models, the estimator is estimating the difference between the input and the artifact-free image. From which the estimation of the artifact-free image is easily obtained. The (estimation of an) artifact-free image consists of spatial information, so during inference the SSIM metric is used to obtain the distance between the ground truth artifact-free image and the estimation of the BBDM models.

The BBDM models in the experiments are by design conditioned on the artifact affected image via the Brownian Bridge (BBDM). The BBDM model is trained with maximal variance parameter $s = 10, 1, 0.1$ percent, referring to these models is done as follows:

- BBDM ($s = 10\%$),
- BBDM ($s = 1\%$),
- BBDM ($s = 0.1\%$).

5.3.1 Synthetic Data Results

For computational reasons, only 50 data points from the test dataset are considered to be able to perform the whole 1000 step sampling procedure.

The overall results can be found in Table 5.2, supported by multiple figures in Appendix A. Figure A.7 to A.9, show a boxplot for each BBDM model, showcasing the effect on the average prediction for 1, 5 and 10 sample predictions. Averaging more samples during inference has a positive influence on all BBDM models.

Overall, it is obvious that the BBDM models seem to perform significantly worse than other experimental models or benchmark models.

TABLE 5.2: Performance (SSIM) of the BBDM models, trained with maximal variance parameter equal to 10, 1 and 0.1 percent. The performance was measured on 50 data points from the test partition. Each model's performance is measured on the average of 1, 5 and 10 drawn samples during inference. In the first column the trained models are listed, in the second column the number of drawn samples are shown with the corresponding performance in the final column.

Model	**Samples**	**Performance (SSIM)**
BBDM ($s = 10\%$)	1	0.8810
	5	0.9178
	10	0.9260
BBDM ($s = 1\%$)	1	0.9379
	5	0.9617
	10	0.9656
BBDM ($s = 0.1\%$)	1	0.9710
	5	0.9838
	10	0.9856
UNET ($df = 100\%$)	-	**0.9983**

Maximal Variance Parameter (s)

A boxplot for each of the BBDM models with averaging 10 independent sample predictions can be found in Figure 5.3.

Decreasing s increases the performance of the BBDM models substantially. Parameter s maximizes variance in the Brownian Bridge, so this directly means that the BBDM model has less noisy inputs and therefore more information about anatomical structures during inference. Training BBDM with a low s seems desirable, but clinical validation is necessary to continue the experiments with this architecture.

5.3.2 Clinical Data Results

In Figure 5.7 the average predictions of the BBDM models are depicted for 5 patients with a unilateral hip prosthesis.

The predictions from the model with a higher maximal variance parameter s lose a lot of anatomical structures which are clearly visible in the input image. Only the shape of the metal and some bone structures are present in the prediction. Clearly, only the information of high contrast edges withstand the addition of noise during the Brownian bridge process. When reducing the maximal noise in the Brownian bridge by reducing s, more details become visible in the prediction.

However, even a maximal variance parameter of $s = 0.1\%$ is still adding enough noise to let the model be very uncertain about the more detailed anatomical structures in the body. This is supported by the stochastic analysis presented in Figure 5.10, where the case $s = 0.1\%$ shows uncertainty about almost all fine details in the image.

5.3.3 Conclusion (BBDM)

The synthetic performance of the BBDM models was already problematic. The clinical results of the BBDM models also lack anatomical consistency with the input image. Even with low s, important details during the Brownian bridge process are lost and therefore the BBDM model is not suitable for clinical applications. The BBDM model was designed to learn a direct mapping between two image domains with diverse image generation, but exact image-to-image translation tasks are too demanding for this generative model.

Removing all stochasticity by setting $s = 0$ will solve the problem of losing information, but this will reduce the BBDM architecture to just $T = 1000$ UNET evaluations. Benchmark model DL-MAR already showed good results with only 1 UNET evaluation, so setting $s = 0$ or closer to 0 is not considered as a valuable experiment.

By *superconditioning* the BBDM architecture on the input image at each timestep, the model will have all anatomical information available during the Brownian Bridge process. This is done in exactly the same way as conditioning the DDPM architecture. The results of the Superconditional Brownian Bridge Diffusion Model (SBBDM) experiments will be discussed in the remainder of this chapter.

5.4 SBBDM Experiments

The SBBDM experiments are, analogously to BBDM, evaluated during inference with the SSIM metric to obtain the distance between the ground truth artifact-free image and the estimation of the SBBDM models.

To overcome the problematic results of the BBDM experiments, the model is additionally conditioned in each timestep on the artifact affected image, like in the CDDPM

models. This superconditional model is called Superconditional BBDM (SBBDM).

The model is trained with $s = 10, 1, 0.1$ percent maximal variance in the Brownian Bridge to be able to inspect the effect of variance in the model. The models are named accordingly:

- SBBDM ($s = 10\%$)
- SBBDM ($s = 1\%$)
- SBBDM ($s = 0.1\%$)

5.4.1 Synthetic Data Results

For computational reasons, only 50 data points from the test dataset are considered to be able to perform the whole 1000 step sampling procedure.

The overall results can be found in Table 5.3, supported by multiple figures in Appendix A. Figure A.4 to A.6, show a boxplot for each SBBDM model, showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

The performance of the SBBDM models come very close to the best performing model UNET ($df = 100\%$). Varying maximal variance parameter s or the number of samples n is not significantly influencing the performance of SBBDM on the synthetic validation data.

The SBBDM architecture seems to perform a lot better than BBDM. The added value of SBBDM with respect to computationally efficient UNET is yet to be found. Clinical validation is required to find supporting evidence.

TABLE 5.3: Performance (SSIM) of the SBBDM models, trained with maximal variance parameter equal to 10, 1 and 0.1 percent. The performance was measured on 50 data points from the test partition. Each model's performance is measured on the average of 1, 5 and 10 drawn samples during inference. In the first column the trained models are listed, in the second column the number of drawn samples are shown with the corresponding performance in the final column.

Model	Samples	Performance (SSIM)
SBBDM ($s = 10\%$)	1	0.9979
	5	0.9980
	10	0.9980
SBBDM ($s = 1\%$)	1	0.9975
	5	0.9979
	10	0.9979
SBBDM ($s = 0.1\%$)	1	0.9972
	5	0.9978
	10	0.9978
UNET ($df = 100\%$)	-	**0.9983**

5.4.2 Clinical Data Results

The subjective evaluation is done for 5 of the 30 patients in the clinical dataset, by selecting a single slice containing a significant portion of the metal implant for each patient. Then the artifact-free predictions of each SBBDM model is compared with the predictions from the benchmark models O-MAR and DL-MAR.

The objective evaluation is done for all 30 patients in the clinical dataset, by selecting a single slice containing a significant portion of the metal implant for each patient. Then the artifact-free predictions of each SBBDM model and the predictions from the benchmark models are tested on the objective evaluation criteria.

Predictions

The synthetic data results did not show significant improvements for the SBBDM models when taking the average of multiple samples n drawn during inference. Nevertheless, $n = 10$ in the clinical evaluation to be able to consistently compare different experimental models and inspect stochasticity.

The average of these 10 sample predictions is recalled to as the average prediction of the SBBDM models. All separate sample predictions of the diffusion based models can be found in Appendix B for a single image slice of 5 patients with a hip prosthesis.

In Figure 5.4, the (average) predictions of the SBBDM, UNET and benchmark models O-MAR and DL-MAR can be compared for the 5 clinical cases. In Figure 5.8 the average predictions for SBBDM are shown separately.

Subjective Analysis

In the subjective performance analysis, the predictions of the experimental models were analyzed on *hallucinations*, *secondary artifacts* and *remaining artifacts*. Any additional abnormal behavior will also be discussed in this section.

All SBBDM models show no clear sign of hallucinating bone structures or anatomical structures, which is a very good sign towards clinical applicability. However, the thin edges between regions are sometimes more faded than in the input image, which is probably the effect of averaging slight inconsistencies between the multiple samples. Analogously to the CDDPM results, One could argue that the SBBDM models may be hallucinating in and close to the metal implant. However, due to the large disturbance by the metal artifacts close to the metal and the absence of a ground truth, it is not possible to confidently assume how the bone and the implant should be restored underneath the heavy artifacts.

All SBBDM models show no clear sign of secondary artifacts, which is a strict improvement to the commercialized model O-MAR. While in the average predictions of the SBBDM ($s = 1\%$) model no clear secondary artifacts are visible, the individual samples show the introduction of secondary horizontal strike artifacts. The other SBBDM models do not show the same kind of behavior, so this is probably an optimization imperfection. There are different artifacts visible in the different sample predictions of the other SBBDM models, but these look more like remaining artifacts. All SBBDM models show some remaining artifacts, both the bigger shadow cast by the metal implant and the long thinner artifact streaks are considered. The predictions of the models SBBDM ($s = 10\%$) and SBBDM ($s = 0.1\%$) are comparable to CDDPM ($df = 1\%$) and UNET ($df = 1\%$) on remaining artifacts. These four models show the least remaining artifacts and are outperforming both O-MAR and DL-MAR on this criterion.

Overall, the SBBDM models $s = 10, 0.1\%$ perform consistently good, the case $s = 1\%$ performs slightly worse on patient 4 and 5. As mentioned, this could be explained by an optimization imperfection.

Stochastic Analysis

Since multiple samples were drawn during inference for the SBBDM models, one could also calculate standard deviation for each pixel, which could be seen as an uncertainty heat-

map. In Figure 5.11, the average prediction, standard deviation and the average prediction with highlighted standard deviation is shown for 1 clinical case for each SBBDM model. The set of sample predictions for all 5 patients and all experimental diffusion models are included in Appendix B.

In the pixel-wise standard deviation image for the SBBDM models, only artifacts and the negative shape of the metal can be found. Therefore, the models seem to be very capable of identifying the areas in the image which are affected by the metal artifacts. While the models are not so consistent in restoring these affected areas, drawing multiple samples enabled the quantification of their uncertainty. Identifying artifacts and quantifying uncertainty are both very important aspects towards clinical applications.

Analogously remarking that consistently making the same error is still an error and this would not be visible in the pixel wise uncertainty heat map.

Contra-lateral Consistency

The consistency with the input image in the contralateral side to the implant is depicted in Figure 5.12 for all presented UNET, CDDPM, SBBDM and benchmark models.

The SBBDM models show significantly less contralateral consistency. As stated before, it is not straightforward how to interpret this result. With close inspection of Figure 5.8 One can observe that most artifacts are removed and soft tissue regions are smooth. When this is the only reason for the lower contralateral consistency, this could be a positive result for SBBDM models. However, in the earlier sections this chapter, the introduction of secondary horizontal artifacts by the SBBDM ($s = 1\%$) model and smooth thinner edges by all SBBDM models was found.

Together with the significantly lower contralateral consistency presented in this section, this leans towards the conclusion that SBBDM is overcorrecting or miscorrecting in their predictions. This could very well mean that in clinical applications SBBDM models are more prone to removing tiny fractures or other anomalies than the DL-MAR, UNET and DDPM approaches. Further investigation is required.

Image Quality

For each of the 30 patients in the clinical dataset, the same three circular regions of interest are used with respect to the former experiments.

In Figure 5.13 and 5.14 the contrast to noise ratio for bladder minus fat and muscle minus fat were depicted. Also for the SBBDM models, the box plots cover a wide range of values. This shows inconsistent performance for different patients. Therefore, model specific conclusions are limited on the CNR evaluation criterion.

The SBBDM ($s = 0.1\%$) performs comparable to the best performing models CDDPM ($df = 1\%$) and UNET ($df = 1\%$).

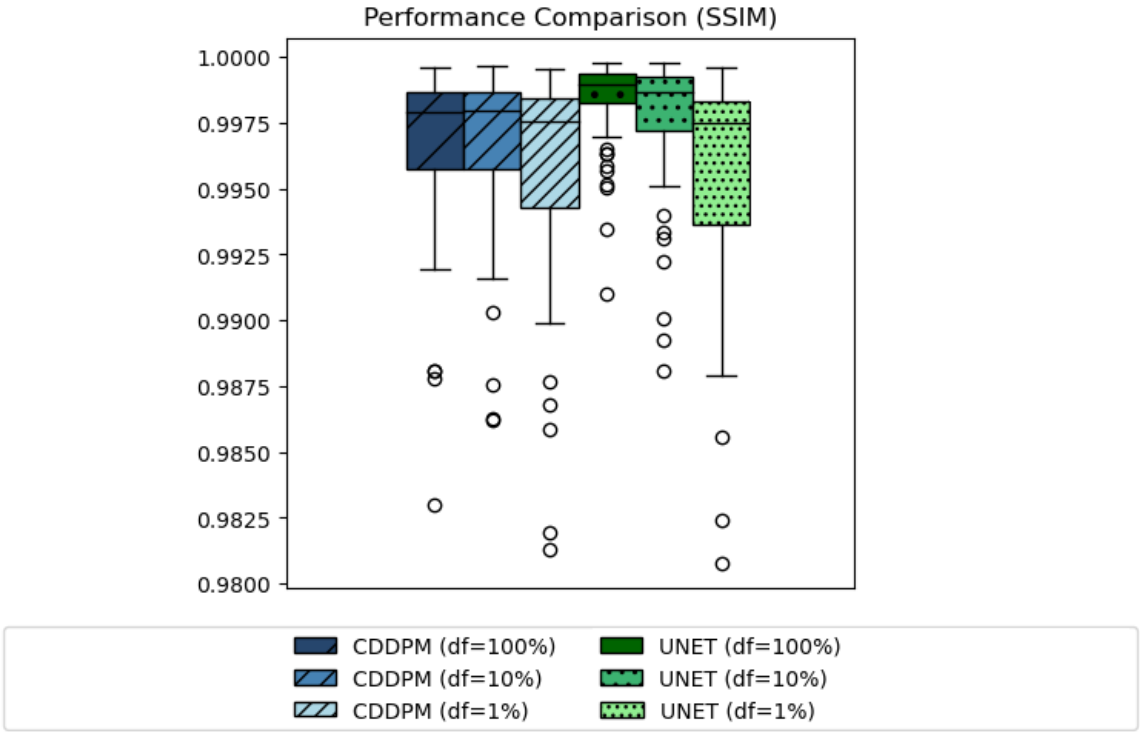

FIGURE 5.1: The performance of the CDDPM and UNET models on 50 data points from the test partition of the simulated dataset, when averaging 10 sample predictions for the CDDPM models.

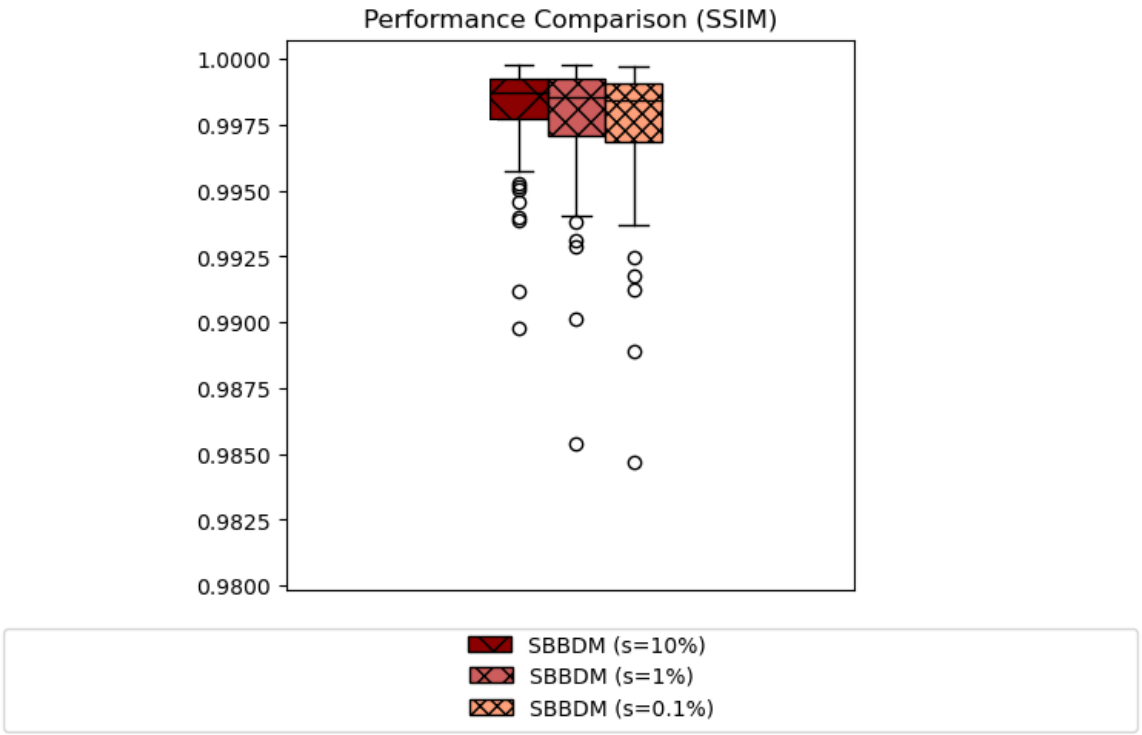

FIGURE 5.2: The performance of the SBBDM models on 50 data points from the test partition of the simulated dataset, when averaging 10 sample predictions. Showcasing the effect of training the model with a maximal variance s of 10, 1 and 0.1%.

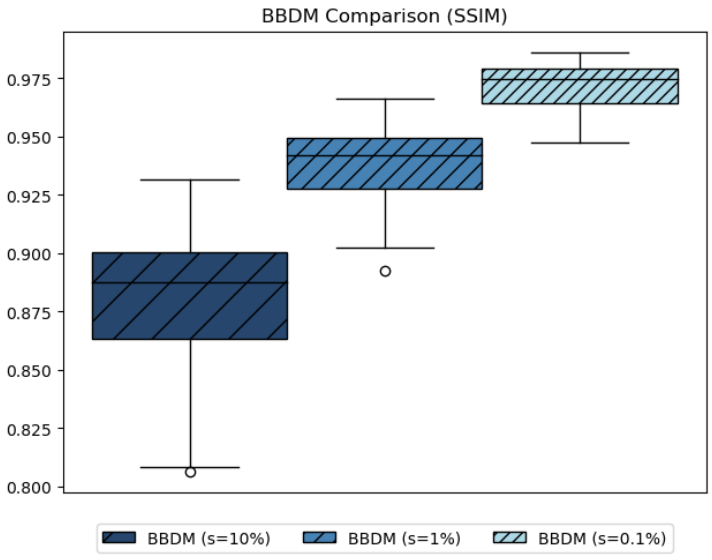

FIGURE 5.3: The performance of the BBDM models on 50 data points from the test partition of the simulated dataset, when averaging 10 sample predictions. Showcasing the effect of training the model with a maximal variance s of 10, 1 and 0.1%.

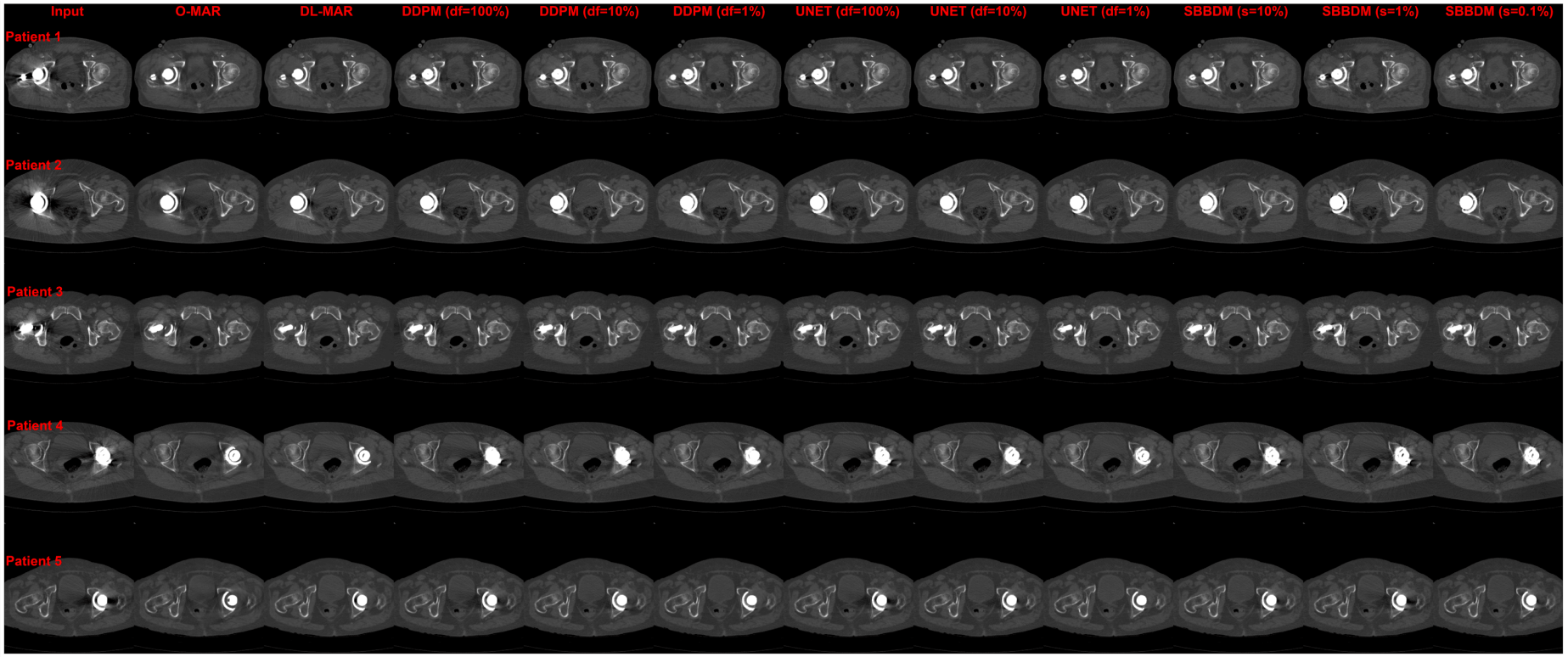

FIGURE 5.4: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the benchmark models (O-MAR and DL-MAR) and the predictions of the experimental models (CDDPM ($df = 100\%$), CDDPM ($df = 10\%$), CDDPM ($df = 1\%$), UNET ($df = 100\%$), UNET ($df = 10\%$), UNET ($df = 1\%$), SBBDM ($s = 10\%$), SBBDM ($s = 1\%$), SBBDM ($s = 0.1\%$)). The artifact affected image is shown in the input column. $[W = 1600, L = 400]$

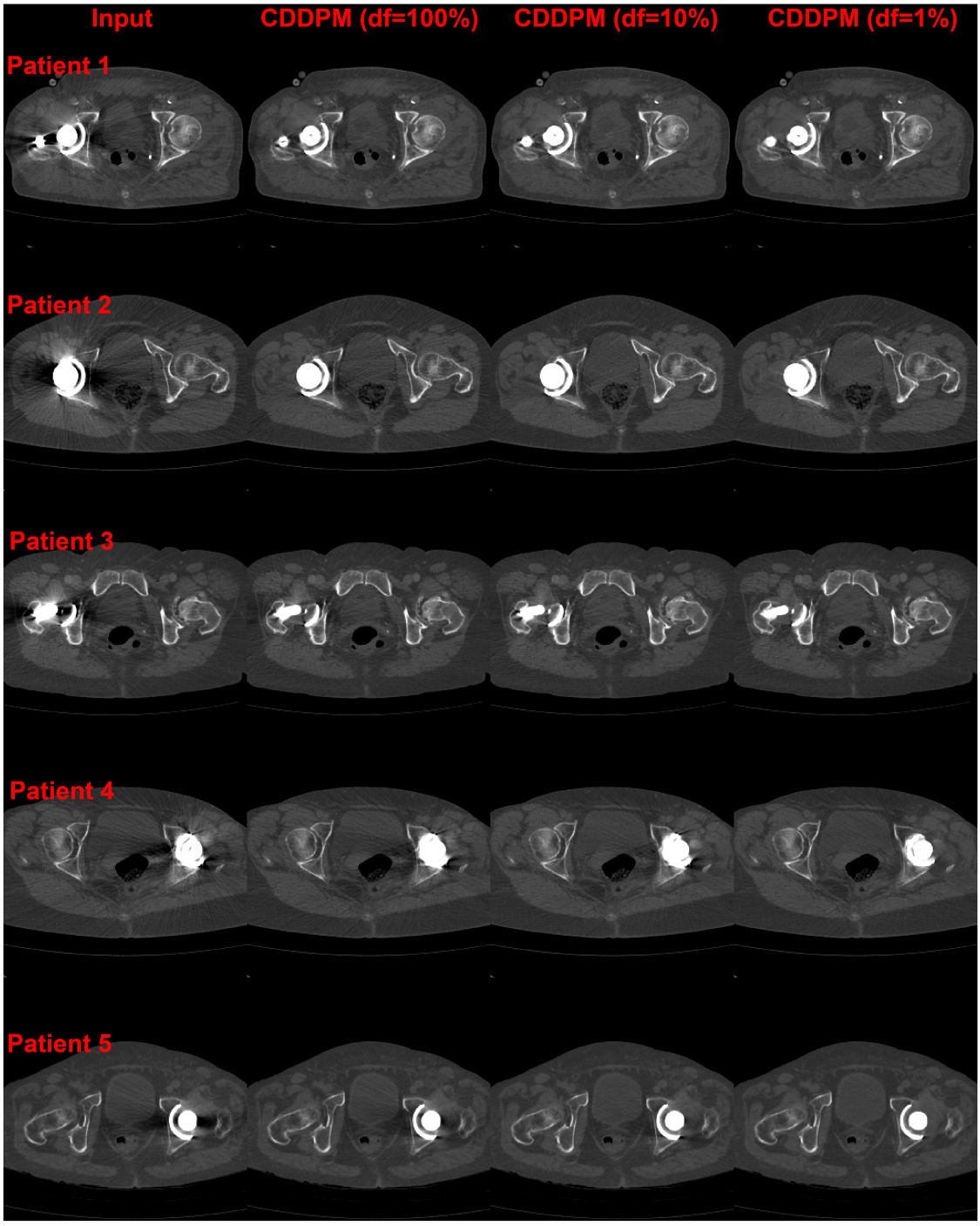

FIGURE 5.5: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the artifact-free prediction of the different CDDPM models, trained on 100, 10 and 1% of the training dataset (df). The prediction of these models is the average prediction of 10 drawn samples. $[W = 1600, L = 400]$

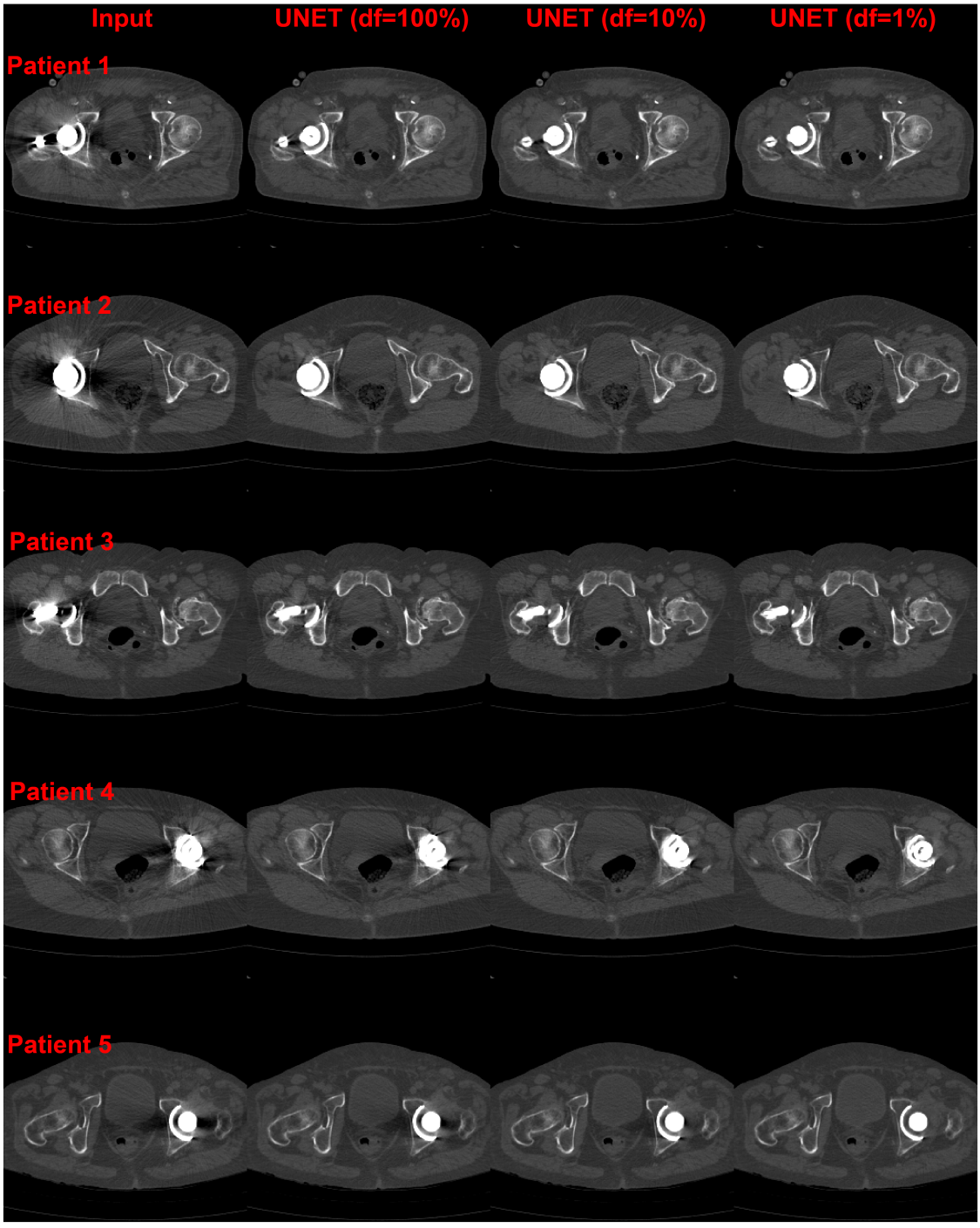

FIGURE 5.6: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the artifact-free prediction of the different UNET models, trained on 100, 10 and 1% of the training dataset (df). $[W = 1600, L = 400]$

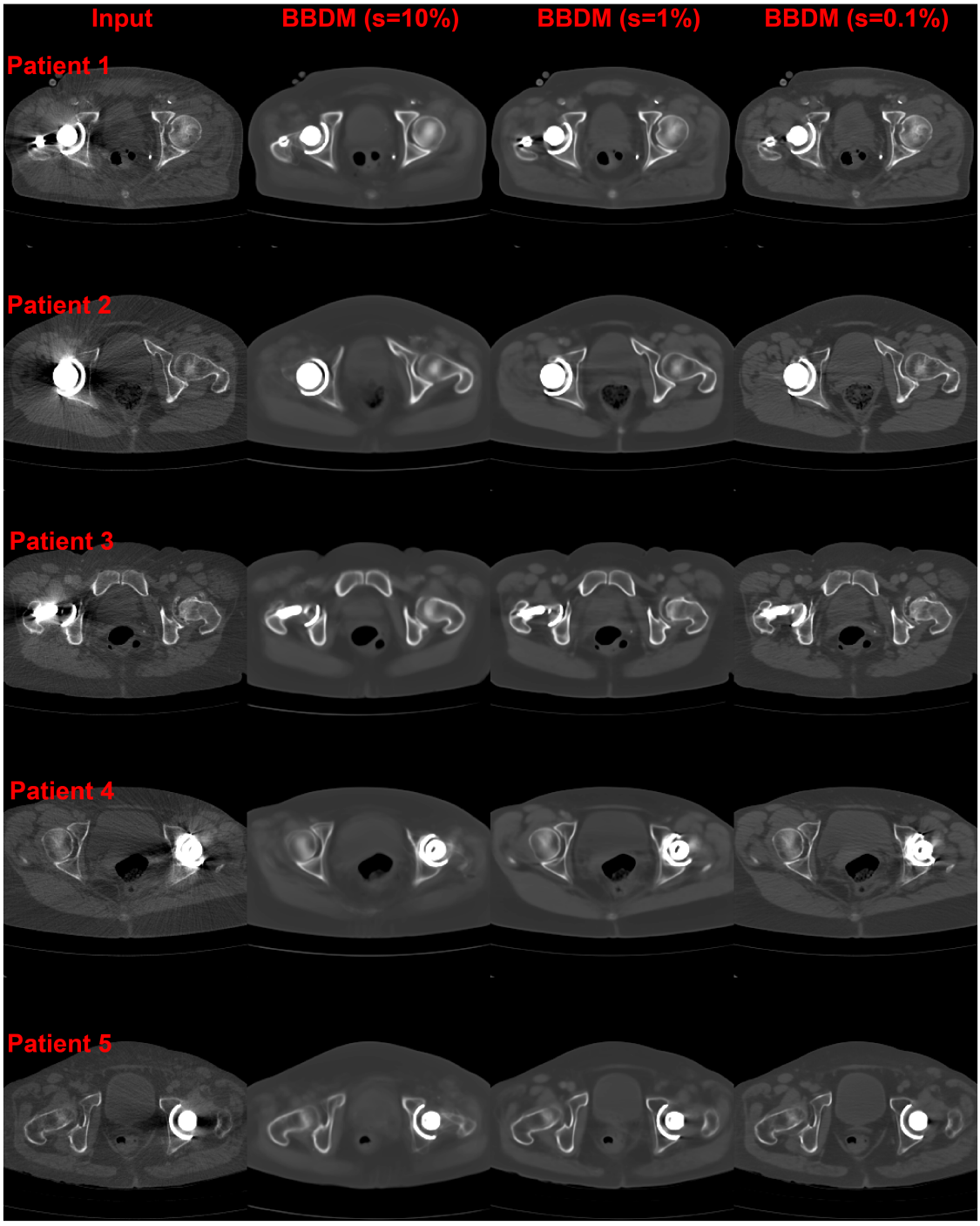

FIGURE 5.7: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the artifact-free prediction of the different BBDM models, trained with a maximal variance of 10, 1 and 0.1% (s). The prediction of these models is the average prediction of 10 drawn samples. $[W = 1600, L = 400]$

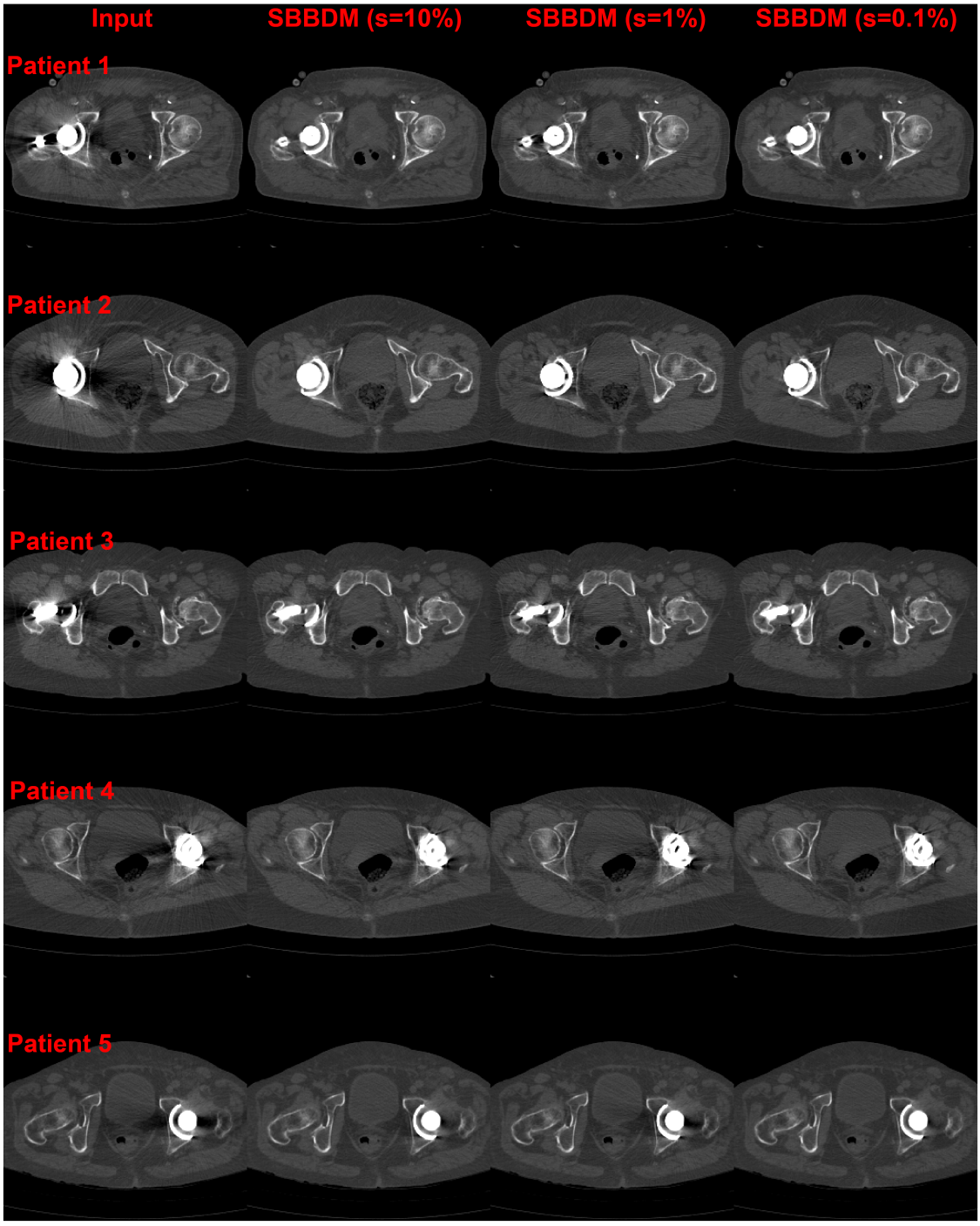

FIGURE 5.8: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the artifact-free prediction of the different SBBDM models, trained with a maximal variance of 10, 1 and 0.1% (s). The prediction of these models is the average prediction of 10 drawn samples. $[W = 1600, L = 400]$

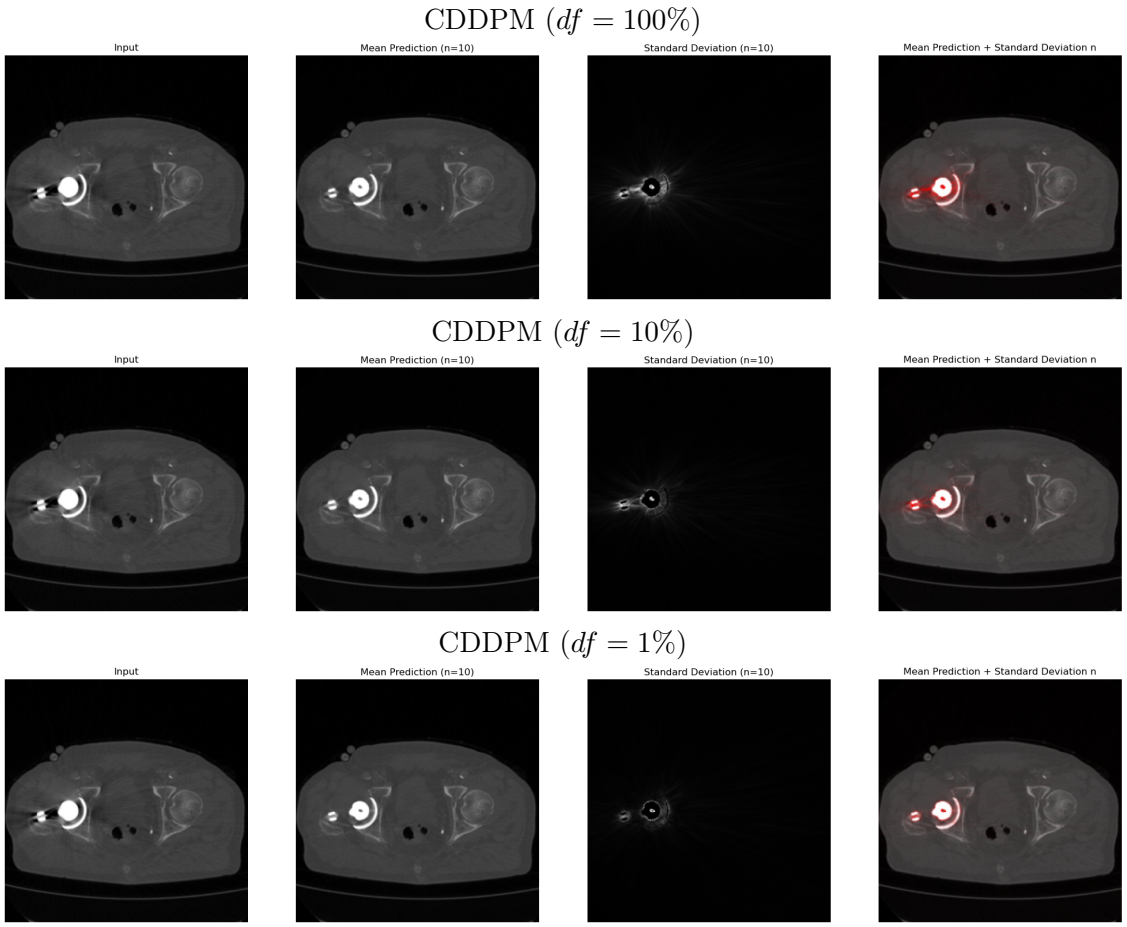

FIGURE 5.9: For each CDDPM model, 1 clinical example is included presenting the input image, the mean prediction, the standard deviation per pixel and the mean prediction with highlighted standard deviation in red. $[W = 4000, L = 1000]$

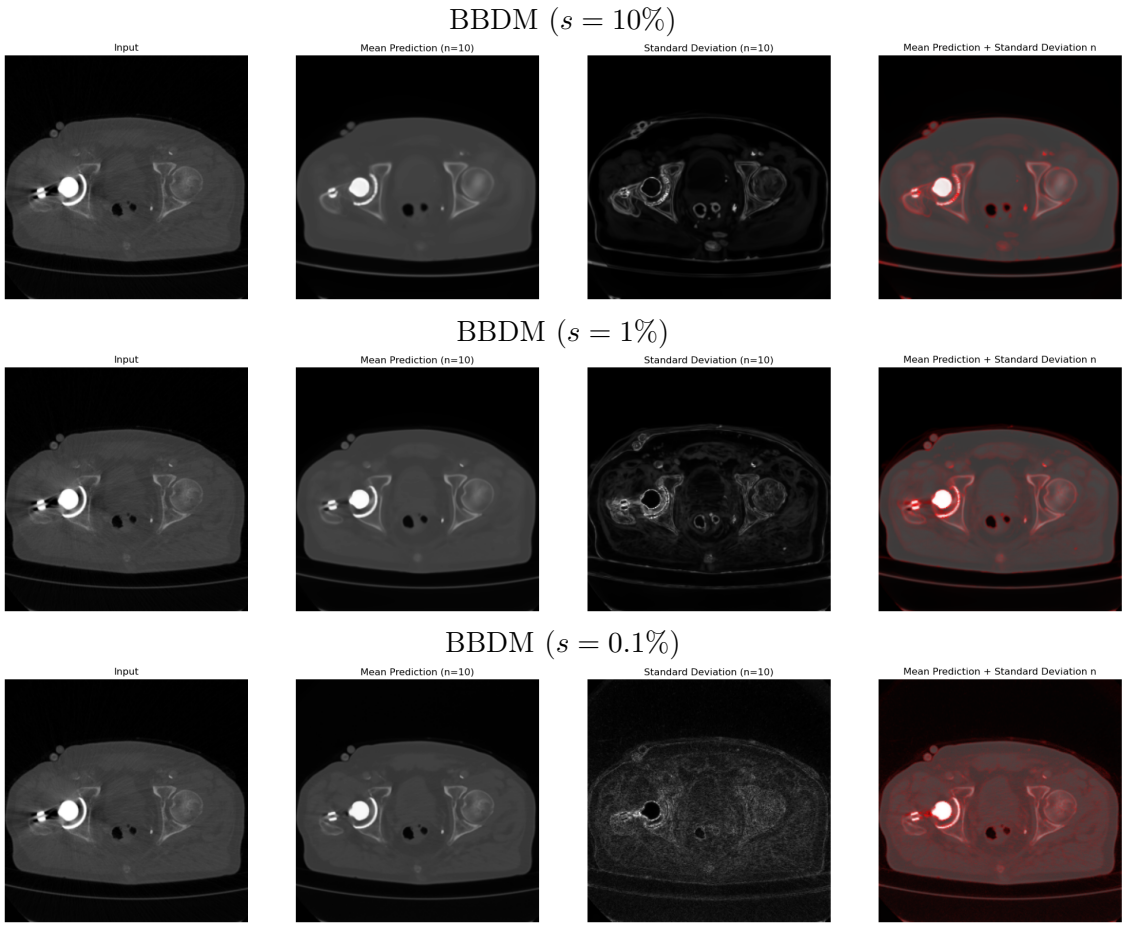

FIGURE 5.10: For each BBDM model, 1 clinical example is included presenting the input image, the mean prediction, the standard deviation per pixel and the mean prediction with highlighted standard deviation in red. $[W = 4000, L = 1000]$

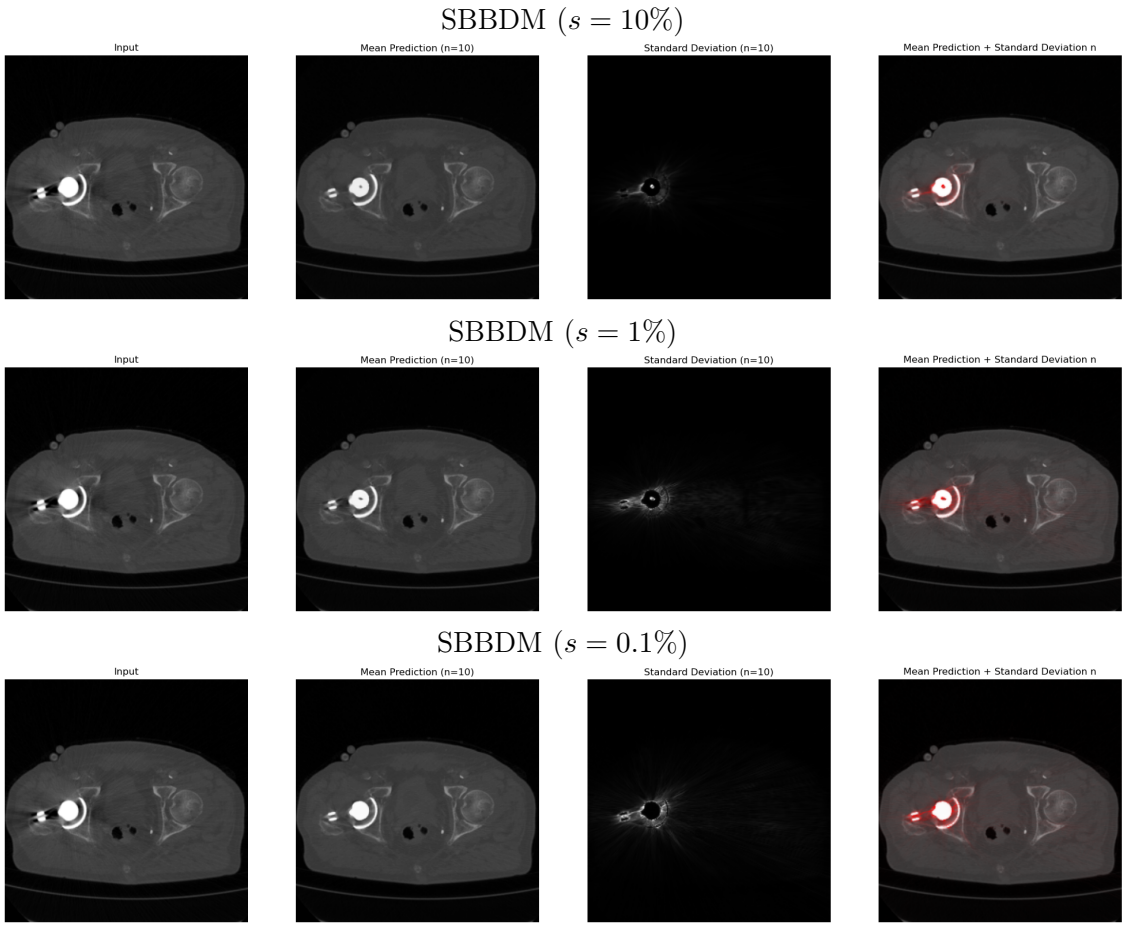

FIGURE 5.11: For each SBBDM model, 1 clinical example is included presenting the input image, the mean prediction, the standard deviation per pixel and the mean prediction with highlighted standard deviation in red. $[W = 4000, L = 1000]$

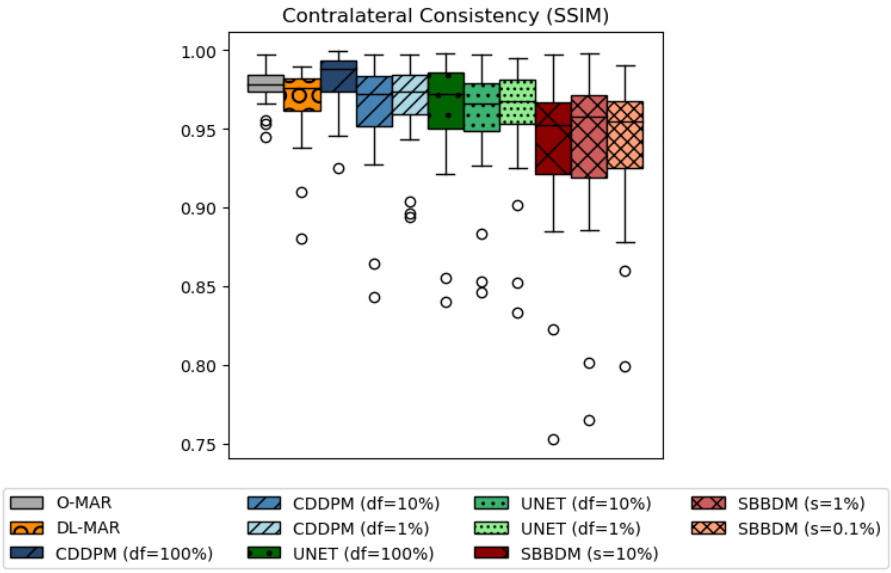

FIGURE 5.12: A boxplot comparison for the evaluation criterion: Contra-lateral Consistency (SSIM). The performance of benchmarkmodels O-MAR and DL-MAR next to the performance of each experimental model based on CDDPM, UNET and SBBDM is shown.

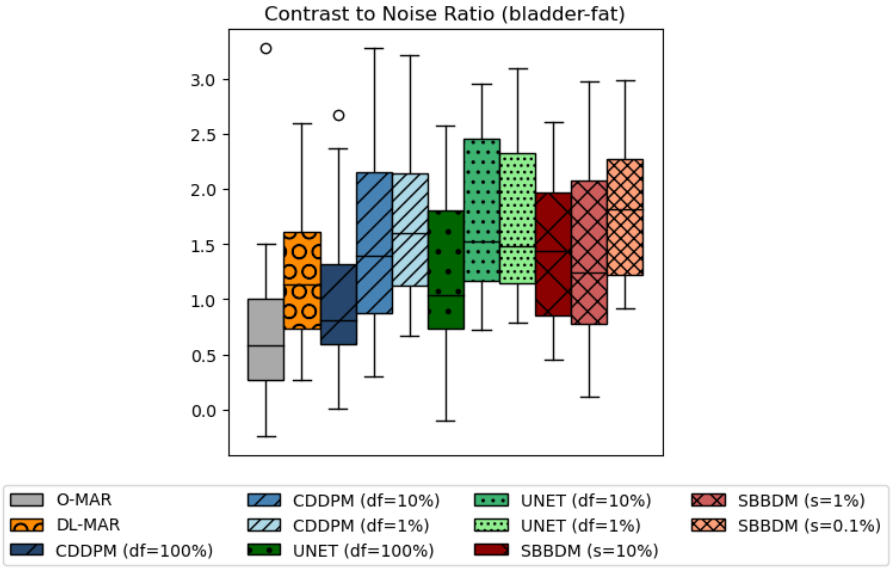

FIGURE 5.13: Boxplot of the contrast to noise ratio in the bladder and fat of the artifact-free predictions from O-MAR, DL-MAR, CDDPM ($df = 100, 10, 1\%$), UNET ($df = 100, 10, 1\%$) and SBBDM ($s = 10, 1, 0.1\%$).

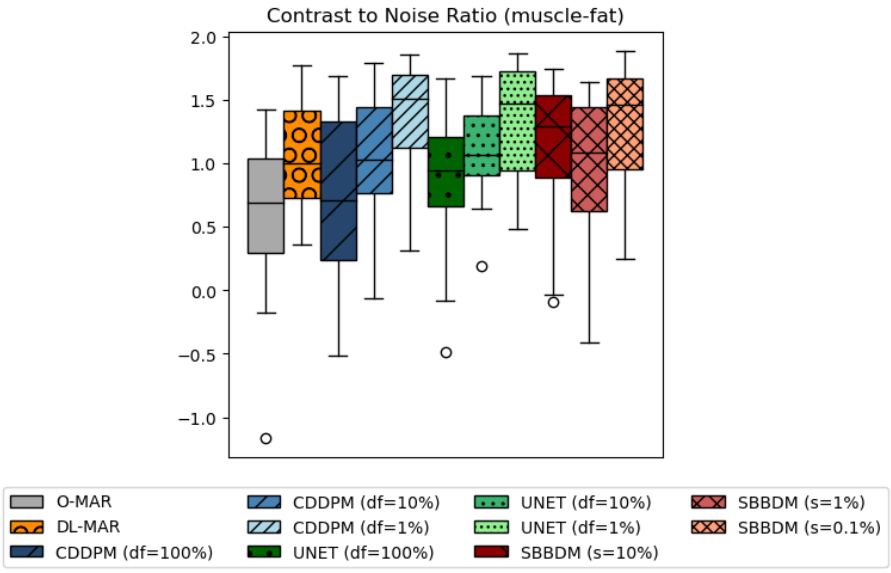

FIGURE 5.14: Boxplot of the signal to noise ratio in the muscle and fat of the artifact-free predictions from O-MAR, DL-MAR, CDDPM ($df = 100, 10, 1\%$), UNET ($df = 100, 10, 1\%$) and SBBDM ($s = 10, 1, 0.1\%$).

Chapter 6

Discussion and Conclusion

This study has explored the application of diffusion-based deep learning methods for reducing metal artifacts in computed tomography (CT) images, specifically Conditional Denoising Probabilistic Models (CDDPM), Brownian Bridge Diffusion Models (BBDM) and their *superconditional* form (SBBDM). The CDDPM model is based on conditioning the pioneering model DDPM by Ho et al. [11] on the input image at each timestep, as done by Saharia et al. [26] and in a medical context by Wolleb et al. [34]. The BBDM model by Li et al. [15] proposed an alternative to the standard conditioning of diffusion models. They proposed to directly map the input domain to the output domain via a Brownian bridge. The SBBDM model is a newly proposed method by this thesis, extending the direct mapping of BBDM with conditioning at each timestep like in CDDPM.

In previous work, diffusion-based models outperformed UNET-based models in various generative tasks and (medical) image-to-image translation tasks, demonstrating enhanced generalizing capacity [1],[11],[26],[27],[34],[35]. Exploring diffusion-based methods for metal artifact reduction (MAR) in CT was a logical progression following the success of the clinically validated UNET-based DL-MAR method [29].

This research demonstrates that conditional diffusion-based models offer significant advantages over traditional commercial methods like OMAR for metal artifact reduction in computed tomography. However, BBDM proved inadequate for precise image-to-image translation, highlighting the need for anatomically consistent methodologies. Both CDDPM and SBBDM present more promising paths.

Since we could train and clinically evaluate on the same data as Selles et al. for DL-MAR [29], we expected the diffusion-based methods to outperform DL-MAR and advance the clinical applicability of deep learning methods in MAR. However, we found that while the diffusion models can offer promising performance and generalization capabilities, they are computationally exhaustive and do not (significantly) outperform UNET-based methods in this research.

While diffusion models are computationally more exhaustive than UNET-based approaches like DL-MAR, diffusion based models did not (significantly) outperform UNET models during synthetic and clinical validation. An ablation study using the same UNET model that was used as a backbone for the denoising in the diffusion models showed that it slightly outperformed the diffusion-based models itself. Moreover, the UNET model outperformed DL-MAR in certain scenarios, indicating that further refinements could improve DL-MAR's effectiveness on metal artifact reduction.

The fact that diffusion models did not outperform UNET-based methods in this research may not be a result of an inherent limitation of diffusion models applied to MAR. Metal artifact reduction is actually an ill-posed inverse problem, since the corrupted image

could correspond to multiple anatomically plausible artifact-free images. A UNET-based method tries to find a single point estimate, while a diffusion model learns the distribution of the plausible artifact-free image manifold. Therefore, it seems more likely that the problem formulation of this thesis does not leverage the mathematical strengths of diffusion models enough. The conditioning strategy of this thesis, concatenating the artifact-affected image at each timestep to the noisy latent variable, moves away from the generative strengths of diffusion models and moves towards a direct image-to-image translation task. While this conditioning approach provides the denoising backbone of the diffusion models with all available anatomical information, it could very well constraint a powerful generative model to learn a deterministic mapping. For these type of mappings, UNET models are already highly optimized. This research could therefore also serve as an ablation study, highlighting the performance ceiling when a generative model is constrained to solve a problem formulated deterministically.

Recent publications support this claim, other diffusion-based approaches did outperform UNET-based methods. DDPM-MAR [14], proposed by Karageorgos et al., is a DDPM based model that was unconditionally trained to estimate missing sinogram data. First of all, the raw sinogram data does not suffer from reconstruction errors like the reconstructed data (or CT-images) we used in our models. Secondly, DDPM-MAR reformulates the problem as a generative inpainting task, leveraging the strengths of diffusion models. By training an unconditional DDPM on clean sinograms, it learns a powerful prior of desired metal-free projection data. Then DDPM-MAR uses the learned distribution to fill in the regions corrupted by the metal. Therefore, DDPM-MAR could have an advantage with respect to our diffusion models by choice of the denoising domain and by a more suitable formulation of the problem leveraging the strengths of diffusion models.

DiffMAR [3], proposed by Cai et al., utilizes a linear degradation process to simulate the physical phenomenon of metal artifact formation in CT. DiffMAR then learns the simulated linear iterative mapping between the artifact-affected image domain and artifact-free image domain, supported by the integration of Structural Information Extraction (SIE) and Time Latent-variable Adjustment (TLA). In comparison to our conditioning method, DiffMAR actually guides the sampling procedure of the diffusion model in a physics-informed manner. The linear degradation process of DiffMAR resembles the direct mapping of BBDM, leveraging a linear process instead of a Brownian motion to construct the iterative mapping. We have seen that reducing the maximal variance parameter (s) in BBDM, regulating the variance of the noise in the bridge, improved the BBDM predictions drastically, recall Figure 5.7. DiffMAR conditions on extracted structural information from a linearly interpolated prior of the artifact affected image, while we conditioned BBDM on the artifact-affected image at each timestep to propose the model SBBDM. We showed that the predictions of SBBDM were significantly better than the predictions of BBDM. However, this negated the effects of the Brownian bridge, i.e. reducing parameter s did not induce the same drastic improvement in performance as it did for BBDM. The physics-informed guidance of DiffMAR and removing the noise from the mapping would be a promising path to explore with the (S)BBDM framework in further research.

BCDMAR [19], proposed by Luo et al., finds a bi-constrained conditional DDPM method for MAR where the conditioning method is very similar to the conditioning in our CDDPM. BCDMAR also conditions on the denoising UNET in each timestep, but with a pre-corrected image prior to guide the sampling process. Additionallly, BCDMAR constructs a data fidelity term to ensure image consistency. The metal trace is projected with a tissue segmented, timestep dependent, pre-corrected prior image onto the sinogram domain and then via filterd backprojection the artifact-affected image is re-approximated

and compared with the measured artifact-affected CT-image to compute the data fidelity term in each timestep. With this formulation, BCDMAR allows itself to effectively explore the solution manifold rather than approximating a single output. By leveraging more of the generative power of diffusion models, BCDMAR shows state-of-the-art performance and reliable tissue representation around metal regions.

Drawing multiple samples (n) and taking the average from diffusion models was adopted from Wolleb et al. [34]. During experiments in this research, the effects of multi-sampling was explored. Multiple samples enabled stochastic inspection of the models' predictions and the average consistently outperformed single samples during inference. As a remark, a critical issue arises when a diffusion-based model predicts the same miss-correction n times. In such cases, the stochastic interpretation of the results could be misleading. Therefore, it is essential to consider the stochastic results as the confidence of the model instead of the probability that it represents the truth. This nuance limits the added value of the stochastic analysis. Instead of increasing the likelihood of clinical application, in this research the stochastic analysis primarily contributes to interpreting the models‘ behaviour.

Another approach for the stochastic inspection would have been the widely adopted and improved Monte-Carlo dropout [6], where random weights of the network are masked during inference to approximate the Bayesian posterior over the models' parameters. The multiple sample method used in this research performs stochastic inspection without changing any weights during inference. Therefore, the generative trajectory is preserved and a random noise input explores a new trajectory to a solution on the learned solution manifold. In contrast, Monte-Carlo dropout creates local deformations in the model's vector field capturing local parameter uncertainty. In theory, this could lead to predictions outside of the learned solution manifold. A comparison of the multiple sample method and the Monte-Carlo dropout during the experimental phase, would give model specific insight in the practical difference of these methods. The computational costs of the two methods are of the same order, so the multiple sample method suffices for stochastic inspection in the absence of Monte-Carlo dropout. In further research, a combination of the two stochasticity methods could be of added value, where the interaction between the two methods would enable a more comprehensive exploration of the learned distributions' support [39].

6.1 Limitations and Further Research

This research reveals several critical constraints that impact the findings. Key areas of concern include computational efficiency, the available data and the mathematical foundations of the models. Each of these factors presents challenges that may influence the reliability and applicability of the results, highlighting the need for careful consideration in interpreting the outcomes of this research. Moreover, the limitations often spark further research directions, which will be discussed in this section as well with references to recent publications.

6.1.1 Computational Efficiency

The UNET model can be tasked to perform the metal artifact reduction in one evaluation. In diffusion models, a UNET backbone is used to denoise the image gradually, by evaluating the UNET iteratively defined by the number of time steps T. In this research the number of timesteps are set to $T = 1000$, based on the pioneering model DDPM [11]. Therefore, diffusion models are 1000 times slower or computationally exhaustive during inference compared to a UNET model. Drawing multiple samples n during inference improved the

predictions and enabled a pixel-wise stochastic analysis of the experimental diffusion based models. However, this increases the computation time again by a factor n, resulting in 10,000 UNET evaluations.

Overall this is a big limitation during the experimental phase, consuming a considerable amount of time and resources compared to a single UNET evaluation. The applicability of the trained diffusion models in clinical scenarios is also limited. The costs and computational time of these experimental diffusion models will be a large factor, especially when a diffusion model is applied after a CT scan and all image slices with metal artifacts are considered instead of a single image slice.

The model DDIM [30], proposed by Song et al., followed DDPM [11] with an implicit method to sample during inference much more efficiently, without changing the training algorithm. Recent works, Timestep Tuner [36] by Xia et al. and AutoDiffusion [16] by Li et al., also propose more efficient timestep methods.

Timestep Tuner is a method that can be applied after training, like DDIM. Skipping timesteps often leads to significantly worse predictions, Li et al. argue that this is partly caused by an inaccurate integral direction applied to a timestep interval. With Timestep Tuner a more accurate integral direction can be found at minimal costs, pushing the sampling distribution to the real one. [36]

AutoDiffusion is proposed as a method to simultaneously find optimal timesteps and an optimal architecture of the diffusion backbone in a predefined *unified search space*. In AutoDiffusion, the timesteps and the architecture of the backbone of a pretrained diffusion model are optimized via evolutionary search. [16]

In further research, applying more efficient timestep methods to diffusion models in MAR, could be essential for clinical application. The mentioned timestep optimizers are all post-training, so the tradeoff between accuracy and efficiency when optimizing timesteps in our diffusion models could be quickly explored.

In the case of DDPM, the variance schedule defines how quickly the model denoises the image in each timestep, from complete Gaussian noise to a sample prediction. The linear variance schedule used in this research keeps the images very noisy for a large portion of the timesteps in the backwards diffusion. It seems that, during this first noisy part, valuable resources are wasted on little to no improvements. Different variance schedules can be considered to increase efficiency or performance. In the case of (S)BBDM the maximal variance parameter s was already analyzed, additionally the variance schedule could be tweaked as well. [4]

Transforming the image domain to a latent space, where performing diffusion should be more computationally efficient, could also be a further research direction to increase efficiency. In most applications, the diffusion process is already applied to a lower dimensional latent spaces. [24] However, less dimensional latent representations of the truth will decrease the available information, which should be considered important in a clinical problem where available information is scarce, like in MAR.

6.1.2 Synthetic and Clinical Datasets

To inspect the generalization capacity of diffusion models, the CDDPM and UNET models were trained on 100, 10 and 1% of our training data. Our CDDPM ($df = 1\%$) and UNET ($df = 1\%$) models were trained on 1% of the training data and their respective models CDDPM ($df = 100\%$) and UNET ($df = 100\%$) were trained on 100% of the training data. During the exact evaluation on synthetic test data, the models trained on 100% outperformed the models trained on 1%, as expected. When evaluating CDDPM ($df = 1\%$) and UNET ($df = 1\%$) on our clinical dataset, the models now outperformed

their respective models trained on 100%. It is an open question in this research why a decrease in training data resulted in a better performance during our clinical evaluation.

While the synthetic dataset consists of different metal implants, the clinical dataset only consists of cases with a unilateral hip prosthesis. A hypothesis for the open performance question would be that the 1% of the training data used to train the $df = 1\%$ models consisted of relatively more patients with a hip prosthesis, which would then specialize the $df = 1\%$ models to remove artifacts from hip prosthesis. This would explain the inverted performance of the CDDPM and UNET models during evaluation on our synthetic and clinical datasets. However, the 1% partition of training data was picked randomly and other implants (no hip prostheses) were found in the 1% during inspection of the training data. A full inspection of the balance of different prostheses in the 1, 10 and 100% training datasets is missing in this research. This limits us to take a confident position in this discussion.

As mentioned, a clinical dataset is used in this study to evaluate the performance of our models in real MAR problems. Most studies in the field of MAR lack a clinical evaluation of their models [29], which elevates the findings of this study. Our clinical dataset exclusively contains images of unilateral hip prostheses. These implants are known to cause the most severe metal artifacts in computed tomography (CT) scans, providing a rigorous test for the performance of our diffusion-based machine learning models. However, the hip prosthesis generally has a relatively simple shape in two-dimensional image slices, aside from some transitional regions in the implant. In contrast, other types of implants, such as screws or dental fillings, exhibit finer and more complex details.

Although the models in this research were trained on a variety of implants, including those with intricate details, they were not evaluated on more complex clinical cases other than hip prostheses. Consequently, it was not possible to fully demonstrate the models' capabilities. Additionally, the performance on the clinical dataset may have been hindered by the training on other types of implants. This limitation restricts the comprehensive evaluation of our models on the clinical dataset and leaves the full clinical evaluation for further research.

6.1.3 Mathematical Motivation

The mathematical foundation provided by the authors of BBDM [15] was proven insufficient and incomplete in this research. While the approach to directly map between image domains remains intriguing and interpretable, the results were presented without adequate justification. Consequently, the BBDM gap persists as an unresolved issue, which significantly limits the added value of the theoretical contributions made regarding (S)BBDM.

By exploring the applicability of generative diffusion based models to the specific clinical image-to-image problem (MAR), experimental sidesteps were taken to provide the model with enough information. For instance, the transformation of the models into (super)conditional versions was included. This was done without providing the necessary mathematical arguments to ensure that training effectively converges the models to the desired function. In the theoretical framework of this research, only other experimental evidence was presented as an argument for conditioning. This gap in mathematical proof limits the reliability of the results and their implications for clinical practice.

6.1.4 Evaluation, Application and Optimization

Clearly, all untouched hyperparameters should be analyzed and optimized first. Due to the inefficient nature of our models and our experimental setting, hyperparameter opti-

mization and cross-validation were for example not viable. The models are now applied to metal artifact reduction and clinically evaluated solely on hip prostheses. Completing the clinical evaluation on other implants would obviously be beneficial to clinical application with diffusion based modeling. Moreover, other (clinical) applications of diffusion based models could be explored. If precise image-to-image translation is required and training data is scarce, like cross-modality image-to-image translation in MRI [38], then we would suggest diffusion-based models as a research direction, also motivated by recent publications discussed in this chapter [14], [3], [19]. Evaluating the results of the diffusion-based models on other domains could add interesting insights to this research.

6.2 Conclusion

In conclusion, the current study demonstrates that although diffusion-based approaches achieve state-of-the-art results for metal artifact reduction in computed tomography, they do not surpass competing deep learning-based methods, while requiring significantly more computational resources. Without a principled problem formulation and guiding strategy for sampling diffusion models, the theoretical advantages of diffusion may not be leveraged enough and more efficient deterministic models can prove just as effective. Therefore, the value of diffusion models for metal artifact reduction may be limited and problem dependent.

Bibliography

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 12 2021. URL: http://arxiv.org/abs/2112.00390.

[2] F Edward Boas and Dominik Fleischmann. Ct artifacts: causes and reduction techniques. *Imaging in Medicine*, 4:229–240, 4 2012. URL: http://www.futuremedicine.com/doi/abs/10.2217/iim.12.13, doi:10.2217/iim.12.13.

[3] Tianxiao Cai, Xiang Li, Chenglan Zhong, Wei Tang, and Jixiang Guo. Diffmar: A generalized diffusion model for metal artifact reduction in ct images. *IEEE Journal of Biomedical and Health Informatics*, 28:6712–6724, 11 2024. URL: https://ieeexplore.ieee.org/document/10629037/, doi:10.1109/JBHI.2024.3439729.

[4] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 1 2023. URL: http://arxiv.org/abs/2301.10972.

[5] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 2020. doi:10.1016/j.isprsjprs.2020.01.013.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, pages 1050–1059, 6 2015. URL: http://arxiv.org/abs/1506.02142.

[7] Richard Garnett. A comprehensive review of dual-energy and multi-spectral computed tomography. *Clinical Imaging*, 67:160–169, 11 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S0899707120302928, doi:10.1016/j.clinimag.2020.07.030.

[8] Lars Gjesteby, Bruno De Man, Yannan Jin, Harald Paganetti, Joost Verburg, Drosoula Giantsoudi, and Ge Wang. Metal artifact reduction in ct: Where are we after four decades? *IEEE Access*, 4:5826–5849, 2016. URL: http://ieeexplore.ieee.org/document/7565564/, doi:10.1109/ACCESS.2016.2608621.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in neural information processing systems*, 27, 6 2014. URL: http://arxiv.org/abs/1406.2661.

[10] William A Gray, Shreya Sekaran, James A Tanyi, and John M Holland. Implications of dental artifacts on radiotherapy planning for head and neck cancer. In *Proc. Multidisciplinary Head Neck Symp*, 2012.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 6 2020. URL: http://arxiv.org/abs/2006.11239.

[12] Jiang Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances, Fourth Edition*. SPIE, 8 2022. URL: https://www.spiedigitallibrary.org/ebooks/PM/Computed-Tomography--Principles-Design-Artifacts-and-Recent-Advances-Fourth/eISBN-9781510646889/10.1117/3.2605933, doi:10.1117/3.2605933.

[13] W A Kalender, R Hebel, and J Ebersberger. Reduction of ct artifacts caused by metallic implants. *Radiology*, 164:576–577, 8 1987. URL: http://pubs.rsna.org/doi/10.1148/radiology.164.2.3602406, doi:10.1148/radiology.164.2.3602406.

[14] Grigorios M. Karageorgos, Jiayong Zhang, Nils Peters, Wenjun Xia, Chuang Niu, Harald Paganetti, Ge Wang, and Bruno De Man. A denoising diffusion probabilistic model for metal artifact reduction in ct. *IEEE Transactions on Medical Imaging*, 43:3521–3532, 10 2024. URL: https://ieeexplore.ieee.org/document/10586949/, doi:10.1109/TMI.2024.3416398.

[15] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 5 2022. URL: http://arxiv.org/abs/2205.07680.

[16] Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7105–7114, 9 2023. URL: http://arxiv.org/abs/2309.10438.

[17] Zan Li, Hong Zhang, Zhengzhen Li, and Zuyue Ren. Residual-attention unet++: A nested residual-attention u-net for medical image segmentation. *Applied Sciences*, 12:7149, 7 2022. URL: https://www.mdpi.com/2076-3417/12/14/7149, doi:10.3390/app12147149.

[18] Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. 4 2024. URL: http://arxiv.org/abs/2404.02954.

[19] Mengting Luo, Nan Zhou, Tao Wang, Linchao He, Wang Wang, Hu Chen, Peixi Liao, and Yi Zhang. Bi-constraints diffusion: A conditional diffusion model with degradation guidance for metal artifact reduction. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. URL: https://ieeexplore.ieee.org/document/10638000/, doi:10.1109/TMI.2024.3442950.

[20] Esther Meyer, Rainer Raupach, Michael Lell, Bernhard Schmidt, and Marc Kachelrieß. Normalized metal artifact reduction (nmar) in computed tomography. *Medical Physics*, 37:5482–5493, 10 2010. URL: https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.3484090, doi:10.1118/1.3484090.

[21] Esther Meyer, Rainer Raupach, Michael Lell, Bernhard Schmidt, and Marc Kachelrieß. Frequency split metal artifact reduction (fsmar) in computed tomography. *Medical Physics*, 39:1904–1916, 4 2012. URL: https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.3691902, doi:10.1118/1.3691902.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 4 2018. URL: http://arxiv.org/abs/1804.03999.

[23] Nikolaos Protonotarios. *The Radon transform, its generalization and their applications in PET and SPECT medical imaging*. PhD thesis, (), , , 12 2019. URL: http://didaktorika.gr/eadd/handle/10442/46704, doi:10.12681/eadd/46704.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 12 2021. URL: http://arxiv.org/abs/2112.10752.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 5 2015. URL: http://arxiv.org/abs/1505.04597.

[26] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 11 2021. URL: http://arxiv.org/abs/2111.05826.

[27] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 4 2021. URL: http://arxiv.org/abs/2104.05358.

[28] Mark Selles, Derk J. Slotman, Jochen A.C. van Osch, Ingrid M. Nijholt, Ruud.H.H. Wellenberg, Mario Maas, and Martijn. F. Boomsma. Is ai the way forward for reducing metal artifacts in ct? development of a generic deep learning-based method and initial evaluation in patients with sacroiliac joint implants. *European Journal of Radiology*, 163:110844, 6 2023. URL: https://linkinghub.elsevier.com/retrieve/pii/S0720048X23001584, doi:10.1016/j.ejrad.2023.110844.

[29] Mark Selles, Jochen A.C. van Osch, Mario Maas, Martijn F. Boomsma, and Ruud H.H. Wellenberg. Advances in metal artifact reduction in ct images: A review of traditional and novel metal artifact reduction techniques. *European Journal of Radiology*, 170:111276, 1 2024. URL: https://linkinghub.elsevier.com/retrieve/pii/S0720048X23005909, doi:10.1016/j.ejrad.2023.111276.

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 10 2020. URL: http://arxiv.org/abs/2010.02502.

[31] Fuminari Tatsugami, Toru Higaki, Yuko Nakamura, Yukiko Honda, and Kazuo Awai. Dual-energy ct: minimal essentials for radiologists. *Japanese Journal of Radiology*, 40:547–559, 6 2022. URL: https://link.springer.com/10.1007/s11604-021-01233-2, doi:10.1007/s11604-021-01233-2.

[32] Yi Wang, Yuan-Zhe Li, Qing-Quan Lai, Shu-Ting Li, and Jing Huang. Ru-net: An improved u-net placenta segmentation network based on resnet. *Computer Methods and Programs in Biomedicine*, 227:107206, 12 2022. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169260722005879, doi:10.1016/j.cmpb.2022.107206.

[33] Martin J. Willemink, Mats Persson, Amir Pourmorteza, Norbert J. Pelc, and Dominik Fleischmann. Photon-counting ct: Technical principles and clinical prospects. *Radiology*, 289:293–312, 11 2018. URL: http://pubs.rsna.org/doi/10.1148/radiol.2018172656, doi:10.1148/radiol.2018172656.

[34] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles. *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348, 12 2021. URL: http://arxiv.org/abs/2112.03145.

[35] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *Medical Imaging with Deep Learning*, pages 1623–1639, 11 2022. URL: http://arxiv.org/abs/2211.00611.

[36] Mengfei Xia, Yujun Shen, Changsong Lei, Yu Zhou, Ran Yi, Deli Zhao, Wenping Wang, and Yong jin Liu. Towards more accurate diffusion model acceleration with a timestep aligner. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5736–5745, 10 2023. URL: http://arxiv.org/abs/2310.09469.

[37] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam Harrison, Mohammadhad Bagheri, and Ronald Summers. Deep lesion graphs in the wild: Relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9261–9270, 11 2017. URL: http://arxiv.org/abs/1711.10535.

[38] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. Mri cross-modality image-to-image translation. *Scientific Reports*, 10:3753, 2 2020. URL: https://www.nature.com/articles/s41598-020-60520-6, doi:10.1038/s41598-020-60520-6.

[39] Tal Zeevi, Ravid Shwartz-Ziv, Yann LeCun, Lawrence H. Staib, and John A. Onofrey. Rate-in: Information-driven adaptive dropout rates for improved inference-time uncertainty estimation. *arXiv preprint arXiv:2412.07169*, 12 2024. URL: http://arxiv.org/abs/2412.07169.

[40] Jianxin Zhang, Xiaogang Lv, Hengbo Zhang, and Bin Liu. Aresu-net: Attention residual u-net for brain tumor segmentation. *Symmetry*, 12:721, 5 2020. URL: https://www.mdpi.com/2073-8994/12/5/721, doi:10.3390/sym12050721.

[41] Yanbo Zhang and Hengyong Yu. Convolutional neural network based metal artifact reduction in x-ray computed tomography. *IEEE Transactions on Medical Imaging*, 37:1370–1381, 6 2018. URL: https://ieeexplore.ieee.org/document/8331163/, doi:10.1109/TMI.2018.2823083.

[42] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 11 2015. URL: http://arxiv.org/abs/1511.08861.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 3 2017. URL: http://arxiv.org/abs/1703.10593.

Appendix A

Additional Performance Boxplots

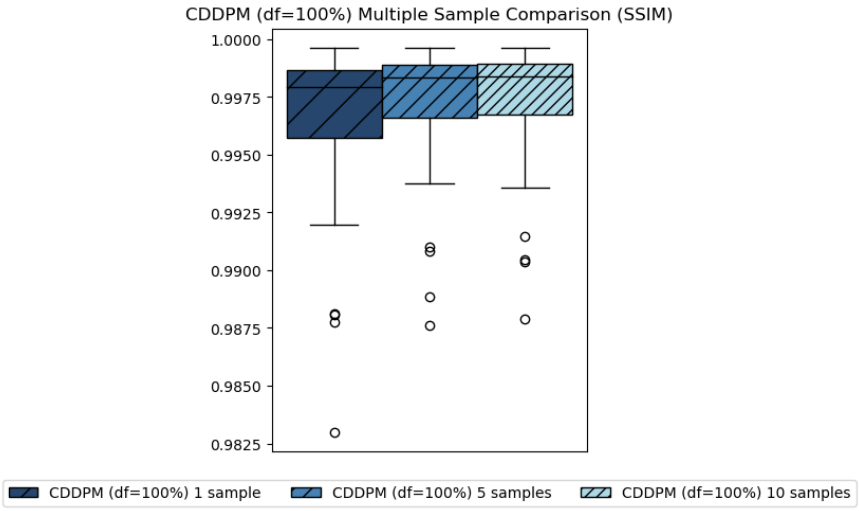

FIGURE A.1: The performance of the model CDDPM (df=100%) on 50 data points of the test dataset, trained on 100% of the training dataset. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

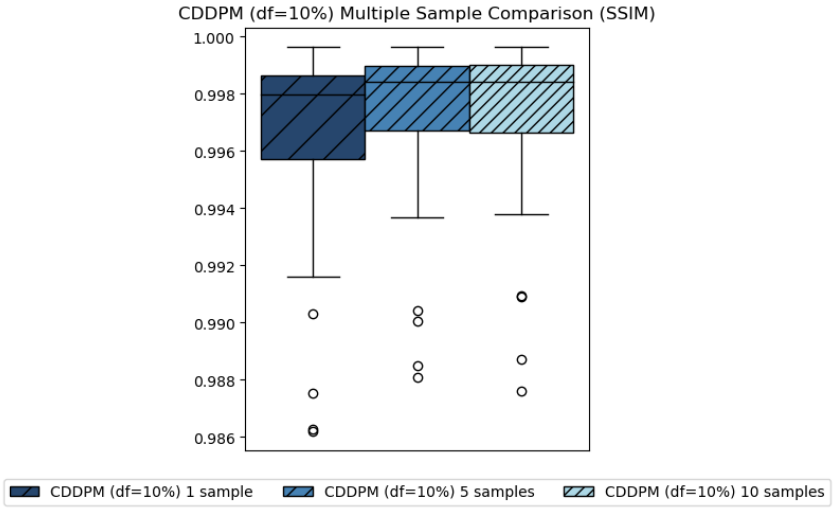

FIGURE A.2: The performance of the model CDDPM (df=10%) on 50 data points of the test dataset, trained on 10% of the training dataset. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

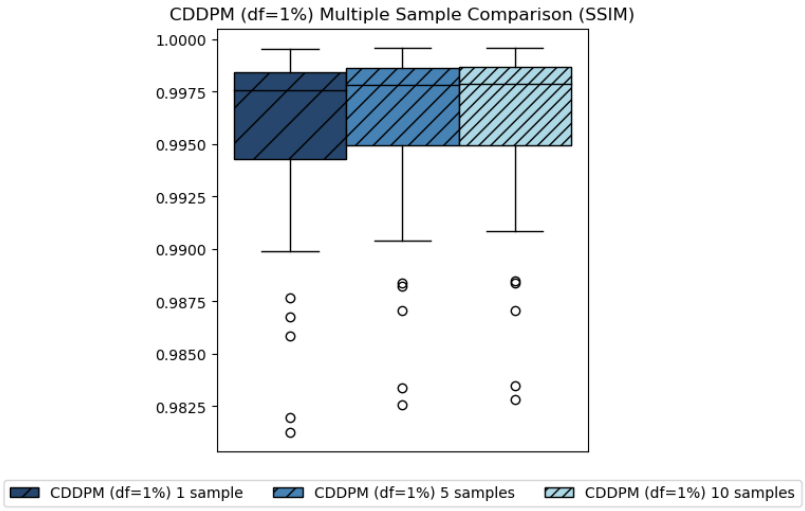

FIGURE A.3: The performance of the model CDDPM (df=1%) on 50 data points of the test dataset, trained on 1% of the training dataset. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

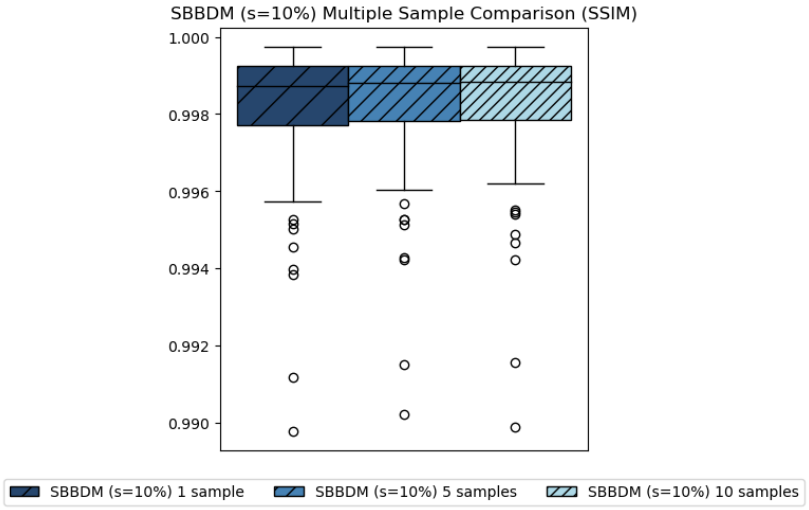

FIGURE A.4: The performance of the model SBBDM (s=10%) on 50 data points of the test dataset, with the maximal variance parameter equal to 10%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

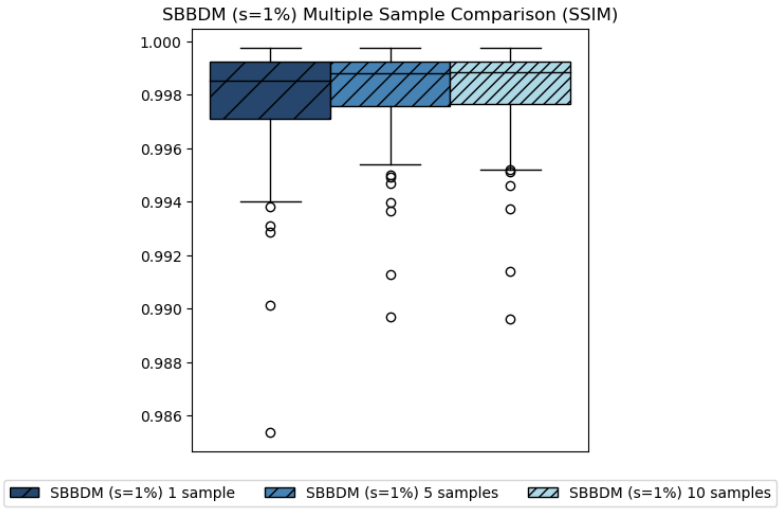

FIGURE A.5: The performance of the model SBBDM (s=1%) on 50 data points of the test dataset, with the maximal variance parameter equal to 1%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

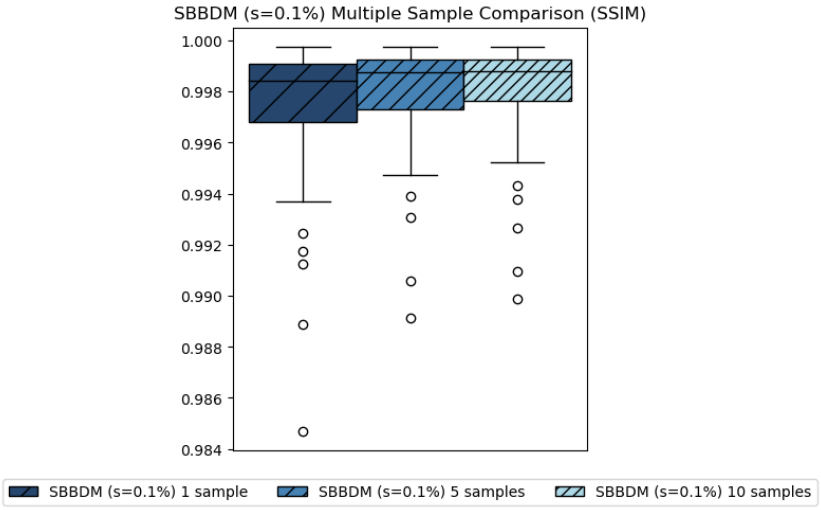

FIGURE A.6: The performance of the model SBBDM (s=0.1%) on 50 data points of the test dataset, with the maximal variance parameter equal to 0.1%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

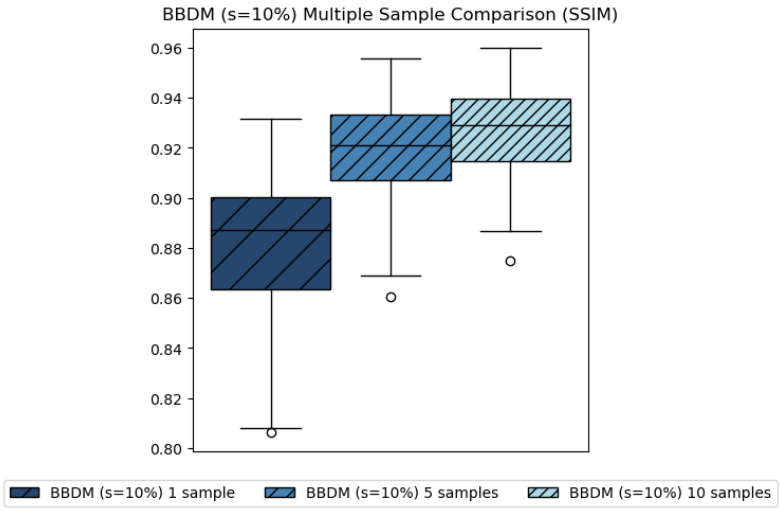

FIGURE A.7: The performance of the model BBDM (s=10%) on 50 data points of the test dataset, with the maximal variance parameter equal to 10%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

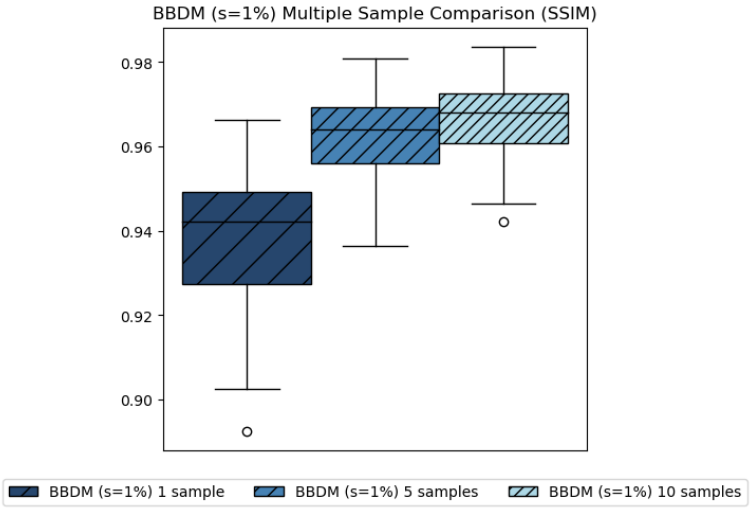

FIGURE A.8: The performance of the model BBDM (s=1%) on 50 data points of the test dataset, with the maximal variance parameter equal to 1%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

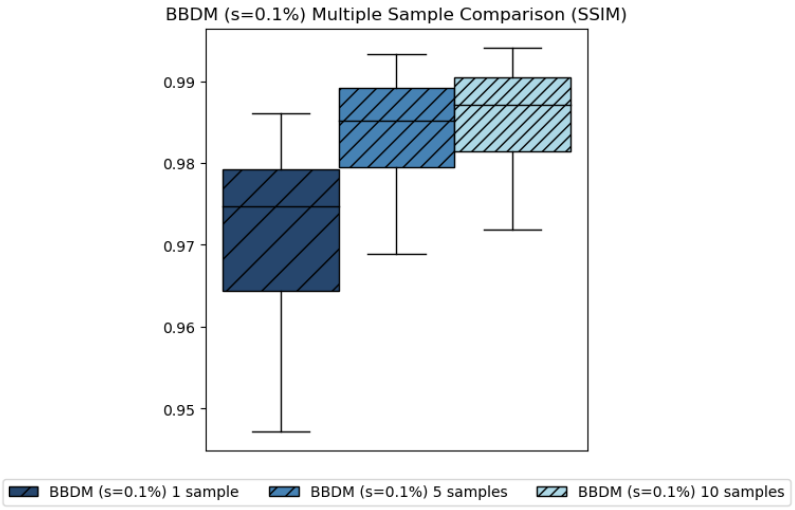

FIGURE A.9: The performance of the model BBDM (s=0.1%) on 50 data points of the test dataset, with the maximal variance parameter equal to 0.1%. Showcasing the effect on the average prediction for 1, 5 and 10 sample predictions.

Appendix B

Additional Clinical Results

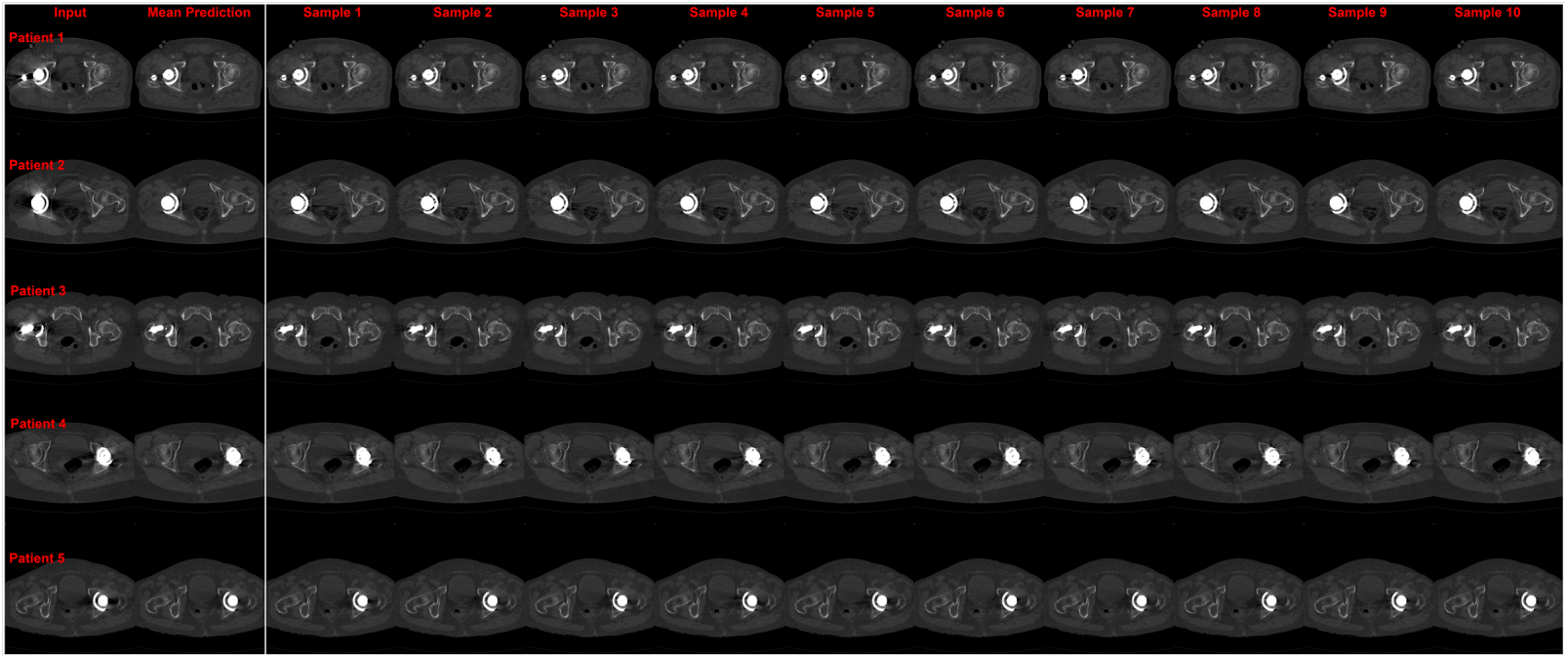

FIGURE B.1: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model CDDPM ($df = 100\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

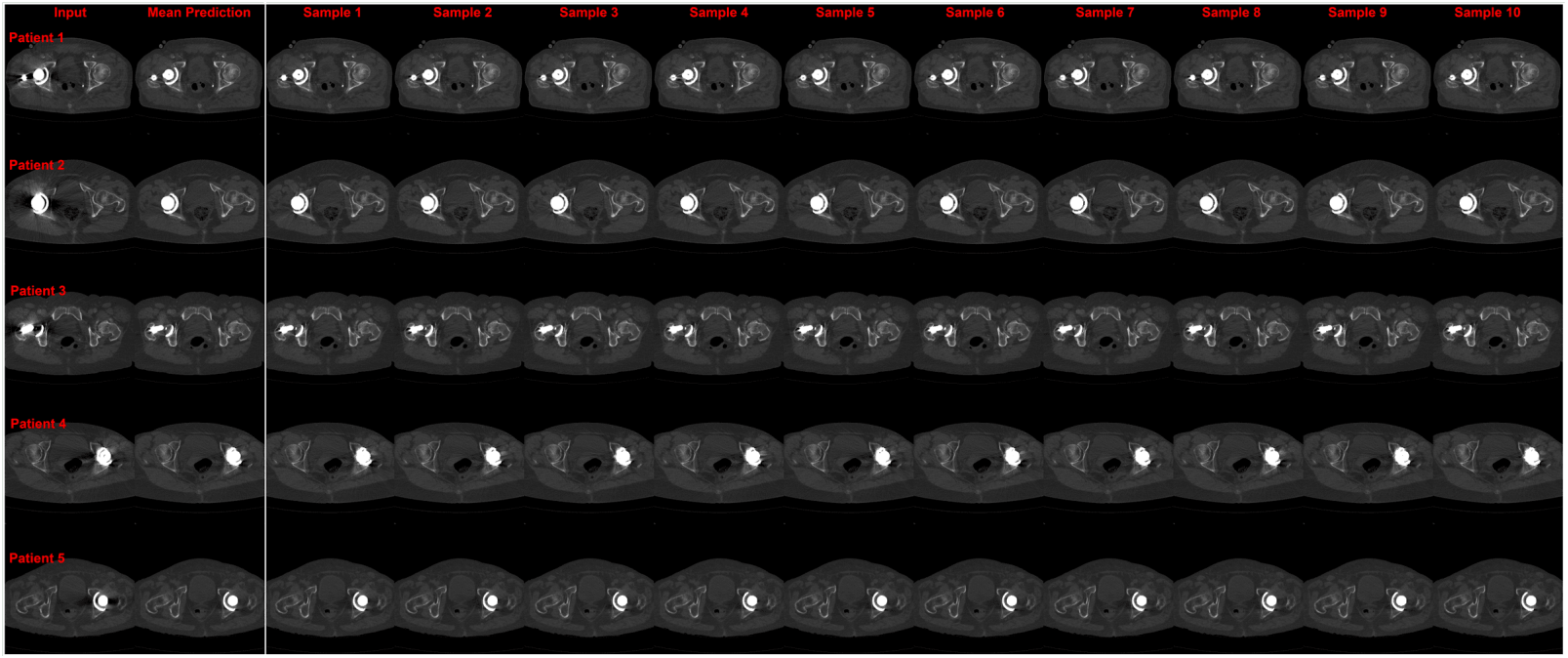

FIGURE B.2: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model CDDPM ($df = 10\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

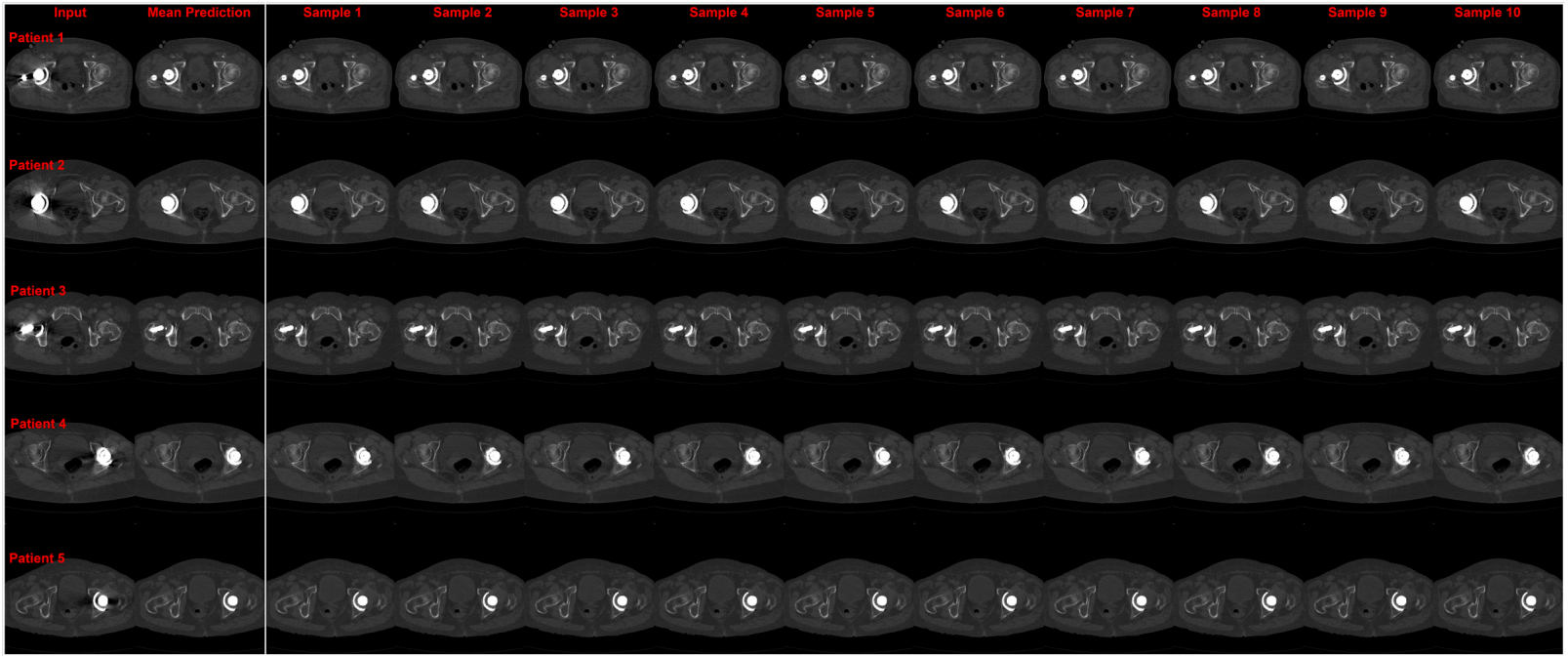

FIGURE B.3: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model CDDPM ($df = 1\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

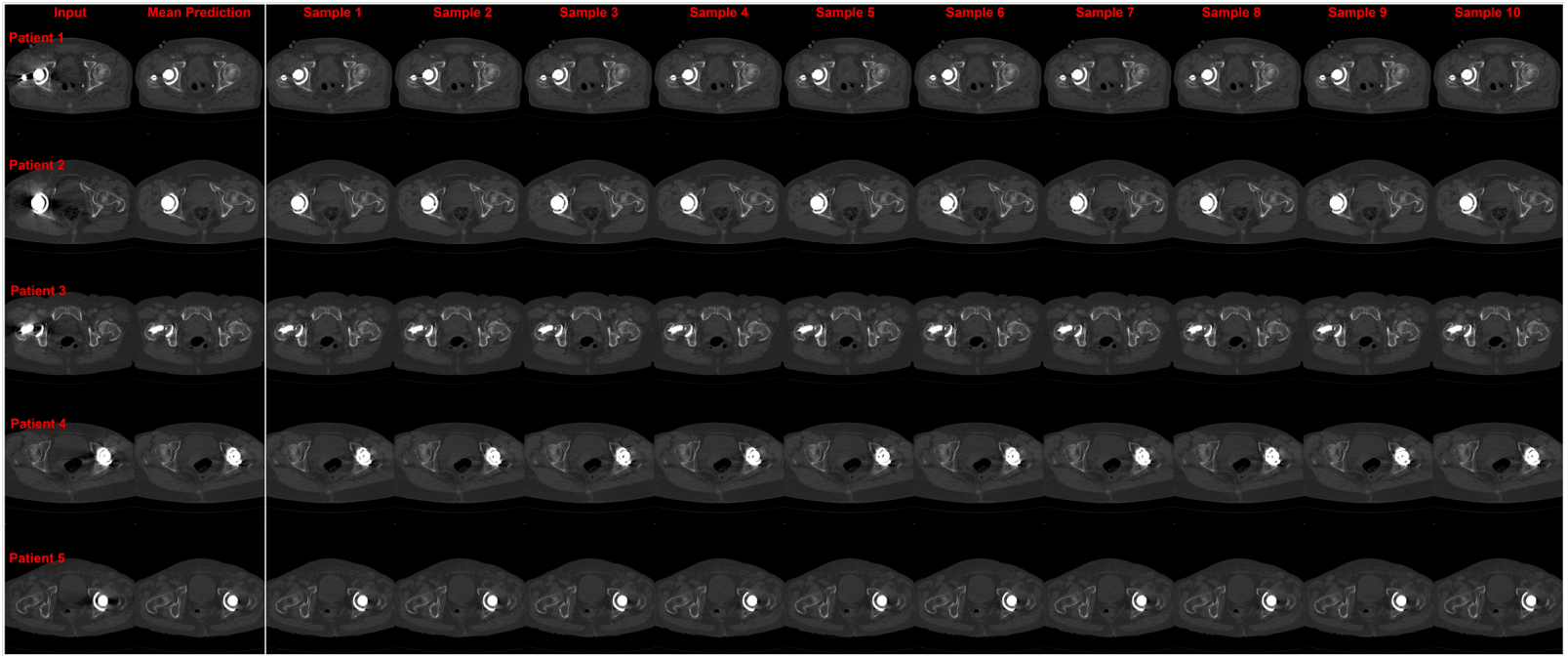

FIGURE B.4: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model SBBDM ($s = 10\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

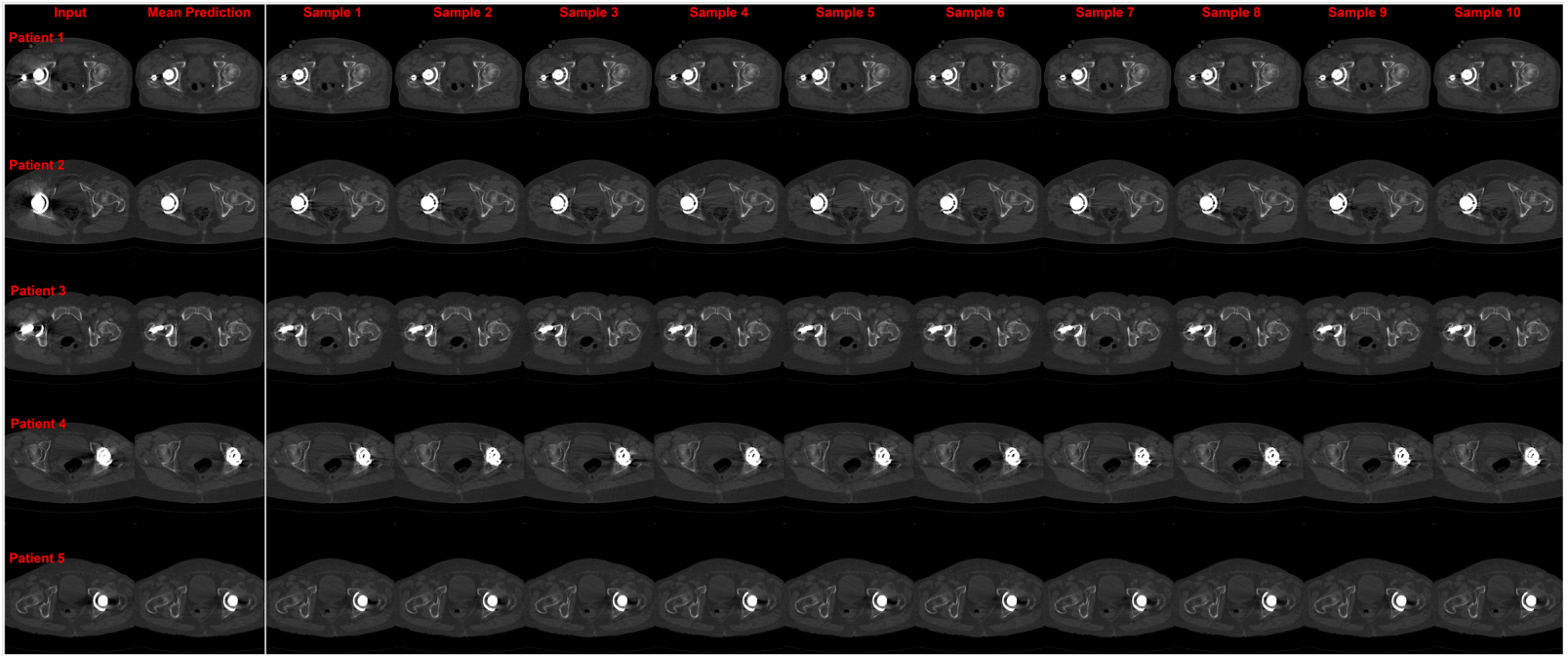

FIGURE B.5: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model SBBDM ($s = 1\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

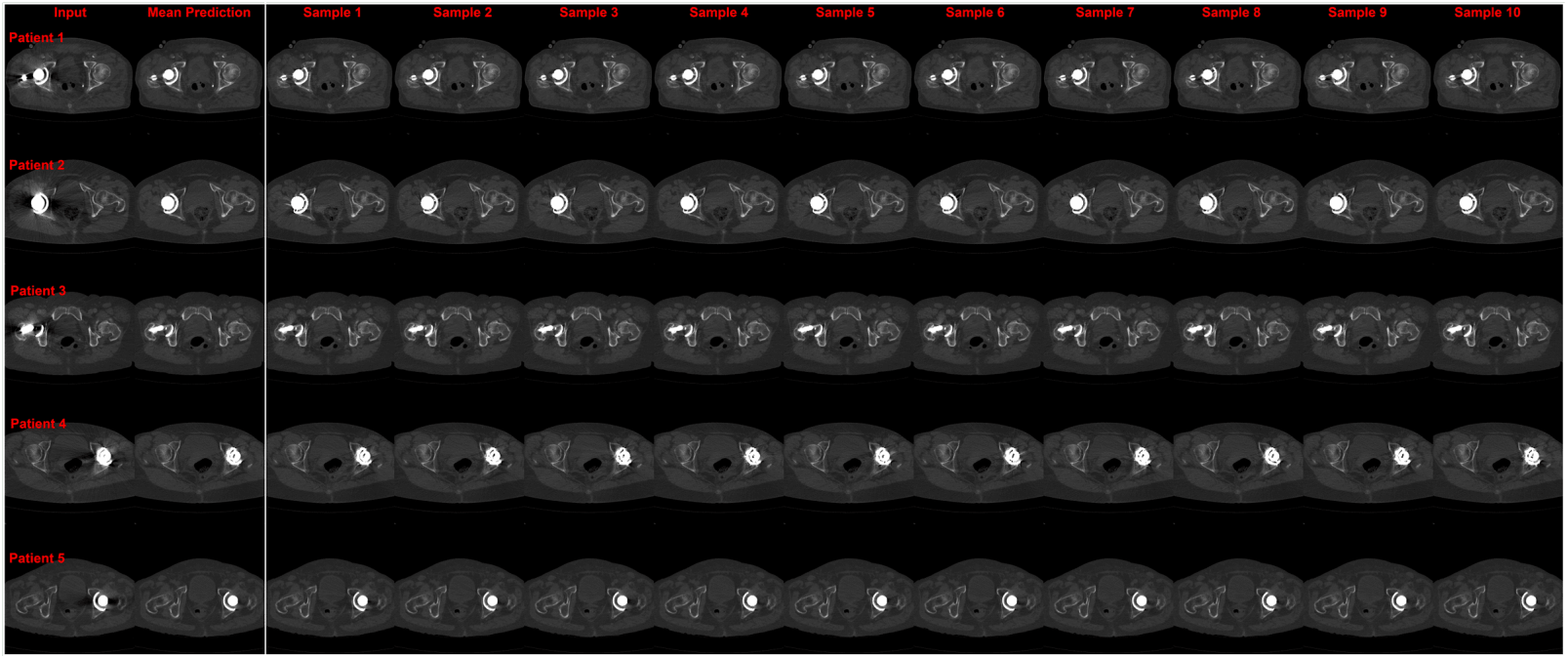

FIGURE B.6: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model SBBDM ($s = 0.1\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

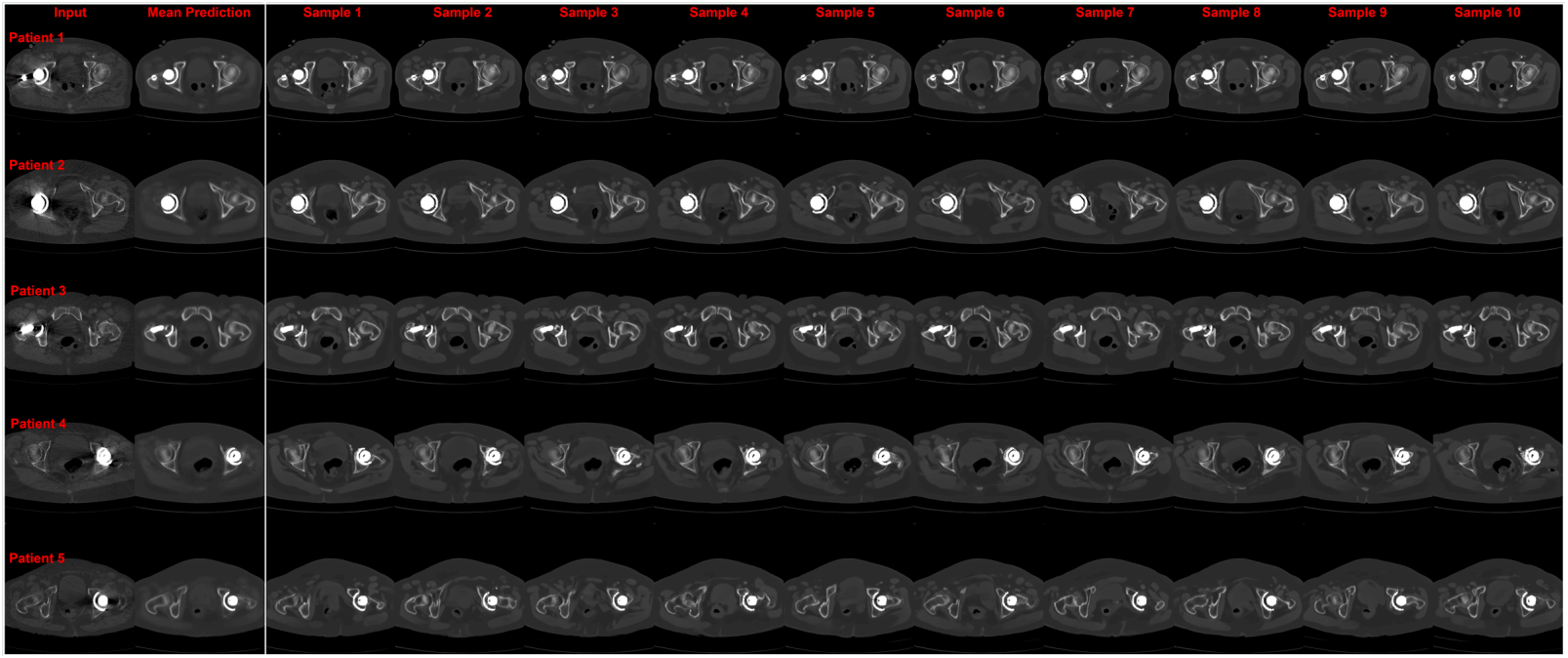

FIGURE B.7: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model BBDM ($s = 10\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

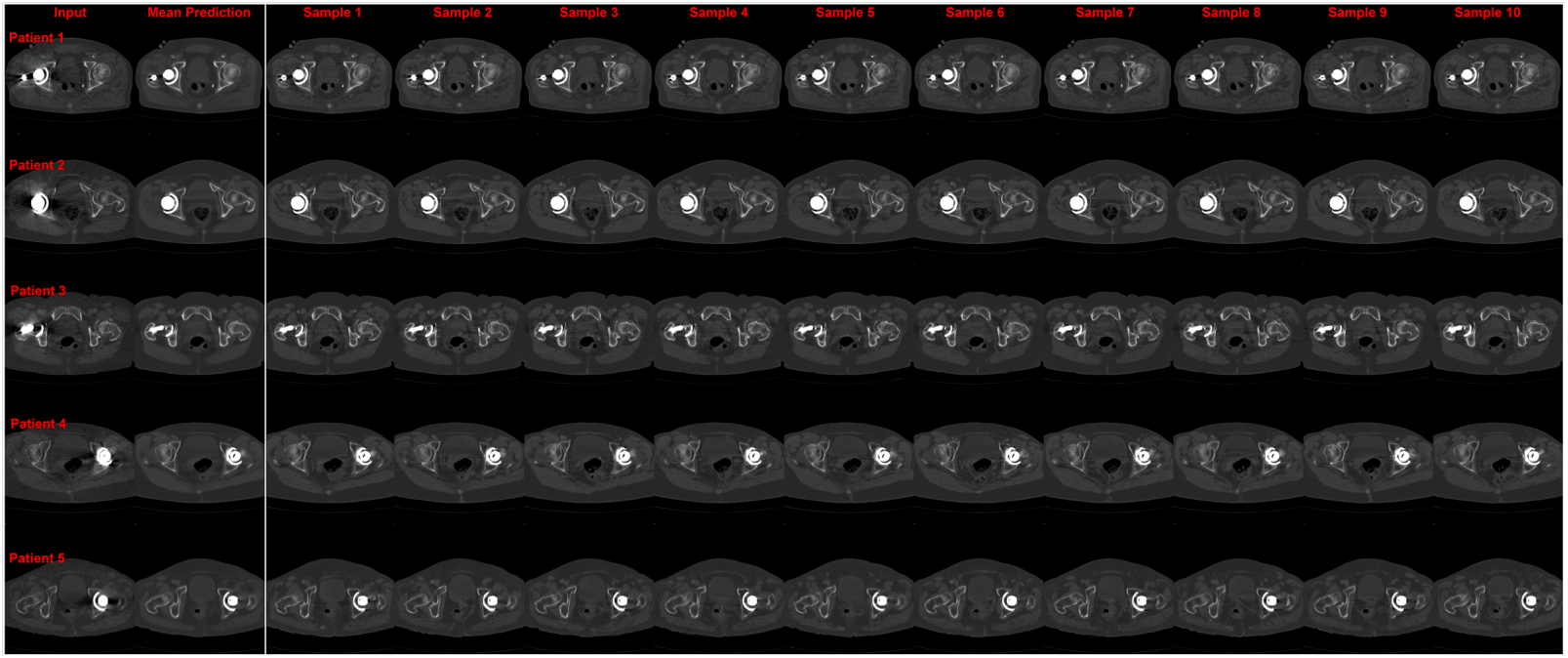

FIGURE B.8: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model BBDM ($s = 1\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$

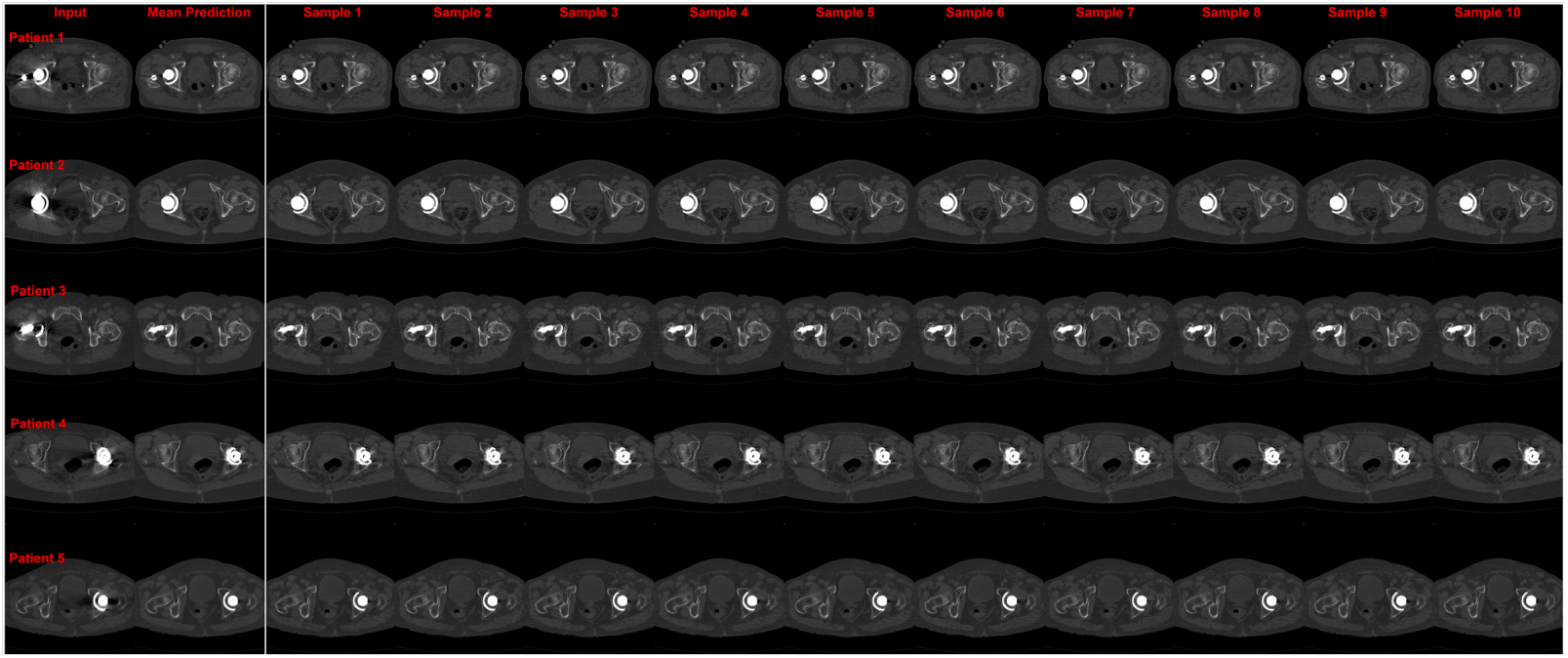

FIGURE B.9: Five clinical cases with severe metal artifacts induced by metal hip prostheses and the predictions of the model BBDM ($s = 0.1\%$). The artifact affected image is shown in the input column, the artifact-free average prediction (n=10) is shown in the mean prediction column and the 10 sample prediction from which the average is constructed are shown in the remaining columns. $[W = 1600, L = 400]$