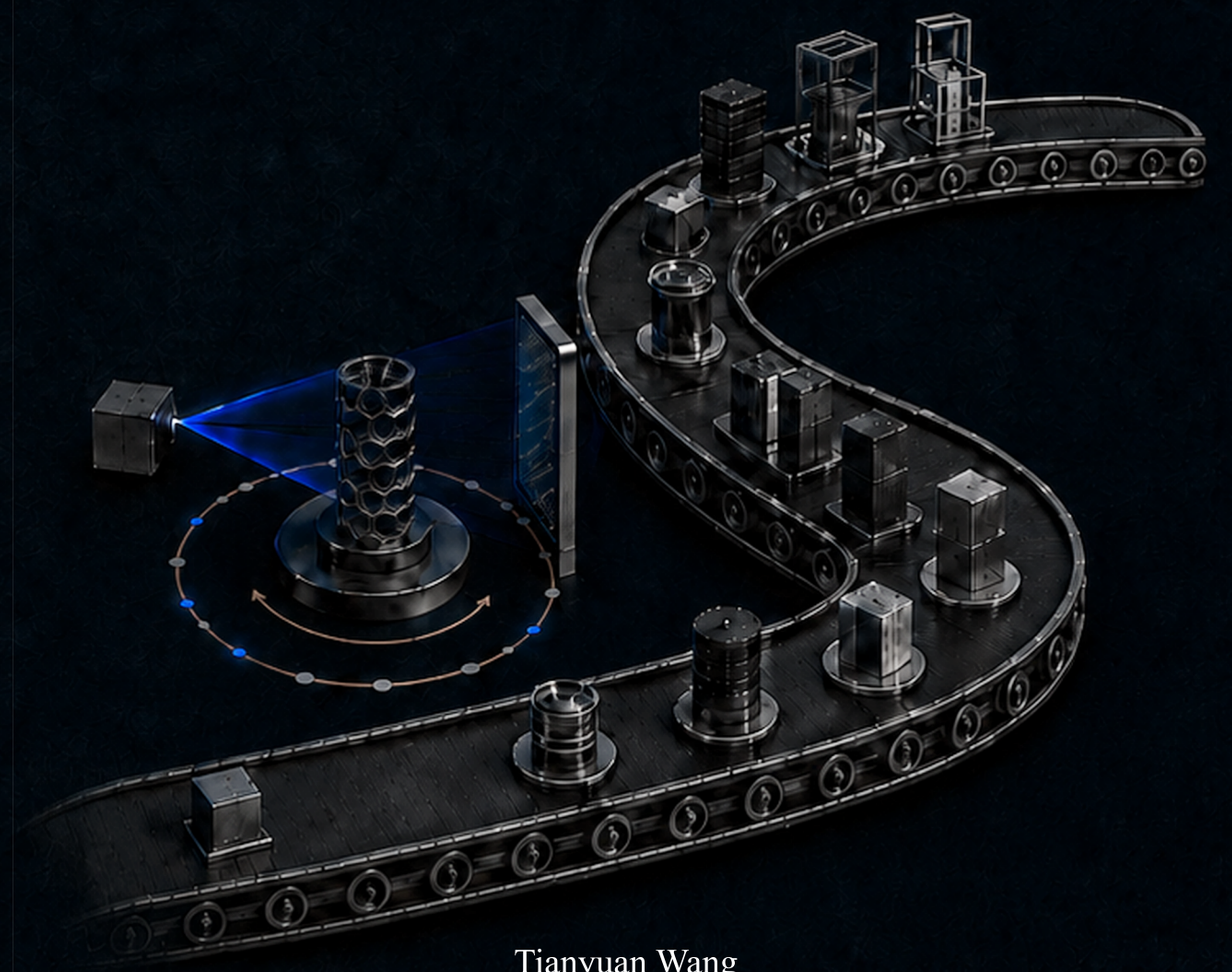
Sequential Optimal Experimental Design for X-ray CT using Reinforcement Learning

Tianyuan Wang

Sequential Optimal Experimental Design for X-ray CT using Reinforcement Learning



Tianyuan Wang

王天元

Sequential Optimal Experimental Design for X-ray CT using Reinforcement Learning

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op woensdag 4 november 2026
klokke 14:30 uur

door
Tianyuan Wang 王天元

geboren te Huaibei 淮北, China

in 1994

Promotores:
Prof. dr. K.J. Batenburg
Prof. dr. T. van Leeuwen (Centrum Wiskunde & Informatica, Universiteit Utrecht)

Co-promotor:
Dr. F. Lucka (Centrum Wiskunde & Informatica)

Promotiecommissie:
Prof. dr. M.M. Bonsangue
Dr. D. M. Pelt
Prof. dr. ir. F.J. Verbeek
Prof. dr. S. Arridge (University College London)
Dr. T. Bubba (University of Ferrara)









The research presented in this dissertation was carried out at the Centrum Wiskunde & Informatica (CWI) in Amsterdam and at the Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands.

This research was co-financed by the European Union H2020-MSCA-ITN-2020 under grant agreement no. 956172 (xCTing).

Copyright © *2026* Tianyuan Wang.

Contents

1 Introduction **1**

1.1 Inline Industrial Inspection 1

1.2 Experimental Design for Inverse Problem Solving 2

1.2.1 Bayesian Perspective on Forward and Inverse Problems 3

1.2.2 Bayesian Optimal Experimental Design 5

1.2.3 Simultaneous vs. Sequential OED 6

1.3 Challenges in Adaptive CT Acquisition 8

1.4 Sequential OED with Reinforcement Learning 9

1.4.1 Formulating Sequential OED as an RL Problem 9

1.5 Research Questions 14

2 Sequential Experimental Design for X-Ray CT Using Deep Reinforcement Learning **21**

2.1 Introduction 21

2.2 Background 25

2.2.1 CT Reconstruction 25

2.2.2 Bayesian OED 25

2.2.3 Reinforcement Learning 26

2.3 Methods 27

2.3.1 Sequential OED as a POMDP 27

2.3.2 Actor-Critic method for policy optimization 29

2.3.3 Network architecture 29

2.4 Numerical Experiments 32

2.4.1 Datasets 32

2.4.2 Implementation 33

2.4.3 Experiment 1 - Uniform informative angles 35

2.4.4 Experiment 2 - Non-uniform informative angles 35
2.4.5 Experiment 3 - Explicit informative angles 38
2.4.6 Experiment 4 - Mixed phantoms with explicit informative angles 41
2.4.7 Experiment 5 - Generalizability Assessment 42
2.4.8 Experiment 6 - Complex phantoms with implicit informative angles 48
2.5 Discussion 49
2.6 Conclusion 55

3 Task-Adaptive Angle Selection for Computed Tomography-Based Defect Detection 57
3.1 Introduction 57
3.2 Background 59
3.2.1 Inverse problem of CT 59
3.2.2 Reinforcement learning for sequential OED 60
3.3 Methods 61
3.3.1 Task-adaptive angle selection for defect detection 61
3.3.2 Reward function 63
3.4 Results 65
3.4.1 Baseline angle selection Methods 65
3.4.2 Shepp-Logan phantoms 65
3.4.3 Simulated industrial dataset 71
3.5 Discussion 83
3.6 Conclusion 84

4 Dynamic Angle Selection in X-Ray CT: A RL Approach to Optimal Stopping 87
4.1 Introduction 87
4.2 Background 89
4.2.1 Forward and Inverse Problems and Optimal Experimental Design (OED) 89
4.2.2 Sequential Optimal Experimental Design (sOED) and Reinforcement Learning 90
4.3 Method 94
4.3.1 Naive optimal stopping 95
4.3.2 Optimal stopping using terminal policy 98
4.4 Results 102

4.4.1 Dataset . . . 102
4.4.2 Implementation . . . 103
4.4.3 Training on synthetic data with Gaussian noise . . . 105
4.4.4 Validation on unseen rotations and noise levels . . . 107
4.4.5 Test on experimental X-ray CT data with two noise levels . . . 110
4.5 Discussion . . . 112
4.6 Conclusion . . . 114
4.7 Data and code availability . . . 114

5 Deep Reinforcement Learning for Optimizing Angle Selection and Dose Allocation in CT Reconstruction 115
5.1 Introduction . . . 115
5.2 Background . . . 118
5.2.1 CT Reconstruction . . . 118
5.2.2 Sequential Optimal Experimental Design via Reinforcement Learning . . . 119
5.2.3 Policy Gradient Methods . . . 120
5.3 Methods . . . 121
5.3.1 Reconstruction with Unequal Dose Allocation . . . 121
5.3.2 Proximal Policy Optimization for CT Environment . . . 124
5.4 Results . . . 127
5.4.1 Datasets . . . 128
5.4.2 Implementation . . . 128
5.4.3 Experiment 1: Reconstruction Using Dose-Aware PnP–PWLS with a TV Prior . . . 130
5.5 Discussion . . . 142
5.6 Conclusion . . . 145

6 Conclusion and Outlook 147
6.1 Contributions and Limitations . . . 147
6.1.1 Contributions and Limitations . . . 147
6.1.2 Outlook . . . 149

A Appendices to Chapter 2 163
A.1 Parameters used in datasets . . . 164

B Appendices to Chapter 4 **165**
B.1 Continuation advantage function 166
B.2 Policy gradient on $\boldsymbol{w}_a$ 168
B.3 Policy gradient on $\boldsymbol{w}_t$ 170
B.4 Additional results 172
B.5 Parameters used in datasets 172
B.6 Architecture of the Actor–Critic network 173

C Appendices to Chapter 5 **175**
C.1 Power Iteration for Lipschitz Constant Estimation 176
C.2 Additional PPO Training Details 176

List of Publications **179**

Summary **181**

Samenvatting **185**

Curriculum Vitae **189**

Acknowledgments **191**

Chapter 1

Introduction

1.1 Inline Industrial Inspection

Non-destructive inspection is essential to ensure that products meet quality standards before further assembly, processing, or distribution. In general, two inspection strategies can be distinguished: *offline* and *inline* [4].

Offline inspection separates the production and inspection stages by extracting samples from the production line. This approach is time-consuming, provides only delayed feedback, and is often not representative since only a limited number of samples are tested [88]. Moreover, offline inspection may damage or compromise the product [4]. For example, transporting parts outside the production environment increases the risk of defects, and in agricultural applications product freshness may be lost during the delay.

In contrast, inline inspection integrates quality control directly into the production process. Products are examined in real time while being transported on conveyor belts or manipulated by robotic arms. This enables immediate feedback and facilitates corrective actions without interrupting production. Inline inspection is therefore highly compatible with automation and aligns with the principles of Industry 4.0 and smart manufacturing [18, 25, 12].

A variety of modalities can be employed for inline industrial inspection, including visible-light cameras, ultrasound, and X-ray imaging [105, 87, 1]. Visible-light cameras cannot access internal features of an object. Ultrasound, on the other hand, is primarily effective for soft materials and homogeneous media; its performance degrades significantly in the presence of air gaps or materials with large acoustic-impedance

mismatches, which cause strong reflections and attenuation. In contrast, X-rays can penetrate a wide range of common packaging and industrial materials, making them far more versatile for internal inspection tasks [24, 83]. However, conventional X-ray radiography provides only two-dimensional projections, where features from different depths overlap and may obscure critical defects.

To overcome these limitations, X-ray Computed Tomography (CT) reconstructs a full three-dimensional representation of an object by acquiring a series of two-dimensional X-ray projections as either the object or the X-ray source–detector pair rotates over 360 degrees. These projections are then combined computationally using tomographic reconstruction algorithms to recover the internal structure of the object with high spatial accuracy, enabling precise localization and quantification of internal features. This capability makes CT highly attractive for inline inspection in domains such as additive manufacturing, food quality control, and agriculture. Figure 1.1 illustrates the concept of inline inspection using X-ray CT, in which products are transported along a conveyor belt. At the inspection station, the conveyor motion is temporarily halted, and a robotic manipulator grips and rotates the product to acquire X-ray projections from multiple viewing angles required for tomographic reconstruction. While CT is already widely used for offline inspection, two-dimensional X-ray imaging remains more common for inline applications [104, 3]. However, deploying CT in an inline setting remains challenging due to the inherent trade-offs between scan time, reconstruction quality, and computational cost. Furthermore, once volumetric data are acquired, automated downstream analysis is required to extract actionable information, such as detecting cracks in industrial components [96], identifying browning in apples [74, 73], or classifying abnormalities in poultry [52]. The efficient integration of CT acquisition with automated analysis is therefore a key requirement for realizing practical inline inspection systems.

1.2 Experimental Design for Inverse Problem Solving

To address the challenges of inline inspection using X-ray CT, it is essential to carefully design the acquisition process. The Figure 1.1 also illustrates the schematic of a CT acquisition setup, consisting of an X-ray source, a detector, and the object in between. *Optimal Experimental design* (OED) refers to the selection of *design parameters* (or control variables), such as projection angles, exposure times, or dose allocation, in

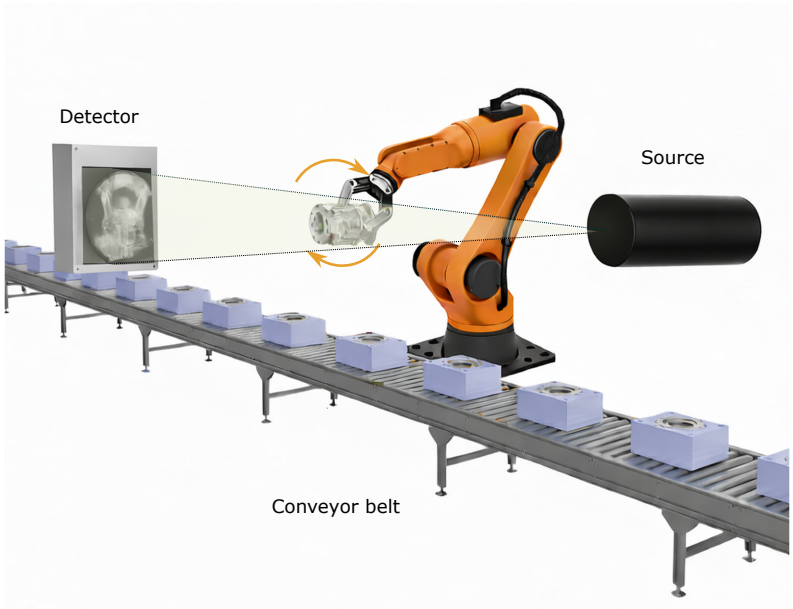


Figure 1.1: Inline inspection using X-ray CT. Products are transported along a conveyor belt to the inspection station. At the scanning position, the conveyor belt is temporarily halted, and a robotic manipulator grips and rotates the product to acquire X-ray projections from multiple viewing angles. This controlled rotation enables projection acquisition while maintaining compatibility with an inline inspection workflow. Illustration created by the author; the inline CT system concept was inspired by [33].

order to optimize the information gained from the measurements. These design choices directly affect the quality of the recorded data and, consequently, the accuracy and reliability of the reconstructed images that form the basis for downstream analysis. In the following, we formulate the problem in terms of projection angle parameters.

1.2.1 Bayesian Perspective on Forward and Inverse Problems

From a mathematical perspective, data acquisition in CT corresponds to a *forward problem*. Once the design parameters are fixed, $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_M\}$, the forward operator $\boldsymbol{A}(\boldsymbol{\theta})$, which represents the discretized Radon transform associated with these parameters, maps the unknown object $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ (with $n \in \mathbb{N}$ denoting the number of pixels or voxels) to a finite set of measurements $\boldsymbol{y}(\boldsymbol{\theta})$ [38, 51]. In practice, this forward process is corrupted by measurement noise, arising from effects such as photon statistics, detector imperfections, or electronic noise. Thus, the acquired data are modeled as noisy measurements. From a Bayesian perspective, the measurement process is

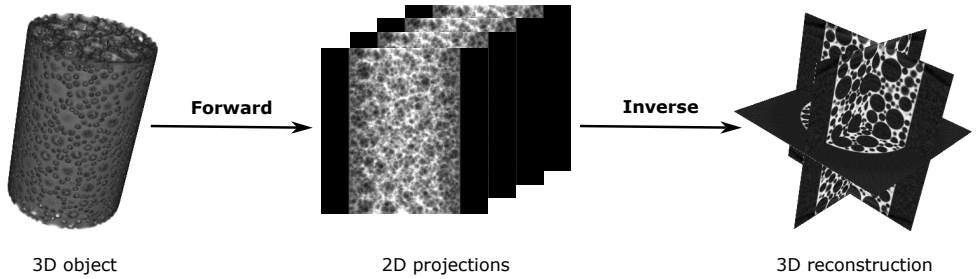


Figure 1.2: Illustration of the forward and inverse processes in X-ray CT using a three-dimensional foam phantom as an example. In the forward process, the CT system acquires a series of two-dimensional projection images as measurements. In the inverse process, these projections are combined computationally to reconstruct the three-dimensional volume of the object. The foam phantom data were obtained from [61]; the projections, reconstruction, and figure composition were generated by the author.

represented by a likelihood distribution:

$$\boldsymbol{y}(\boldsymbol{\theta}) \sim \pi_{\text{data}}(\boldsymbol{y} \mid \bar{\boldsymbol{x}}; \boldsymbol{\theta})\,. \tag{1.1}$$

CT reconstruction can then be formulated as an *inverse problem*: recovering the unknown object $\bar{\boldsymbol{x}}$ from the finite set of noisy measurements $\boldsymbol{y}(\boldsymbol{\theta})$, resulting in an estimate $\widehat{\boldsymbol{x}}$. Because the inversion is ill-posed, direct recovery is unstable and requires the incorporation of prior information. Within a Bayesian framework, this is naturally achieved by combining the likelihood with a prior distribution, yielding the posterior:

$$\begin{aligned} \widehat{\boldsymbol{x}}(\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{x}} \; \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{x}} \; \frac{\pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) \; \pi_{\text{prior}}(\boldsymbol{x})}{\pi(\boldsymbol{y})} \\ &\propto \arg\max_{\boldsymbol{x}} \left\{ -\log(\pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})) - \log(\pi_{\text{prior}}(\boldsymbol{x})) \right\}, \end{aligned} \tag{1.2}$$

where $\pi_{\text{prior}}(\boldsymbol{x})$ encodes assumptions about the image, such as smoothness, sparsity, or structural properties. The maximum a posteriori (MAP) estimate thus balances data fidelity with prior knowledge to produce a stable and reliable reconstruction. Figure 1.2 illustrates the forward and inverse processes in X-ray CT using a three-dimensional foam phantom [61, 41, 79]. In the forward process, the CT system acquires a series of two-dimensional projection images as measurements, while the inverse process reconstructs a three-dimensional volume from these projections. In the two-dimensional projections, internal structures overlap along the projection direction, whereas the re-

constructed three-dimensional volume reveals the true spatial arrangement of internal features. This comparison highlights the necessity of using X-ray CT for industrial inspection, where accurate visualization of internal components is essential.

1.2.2 Bayesian Optimal Experimental Design

In principle, the Optimal Experimental Design (OED) aims to select the design parameters that maximize the information obtained from an experiment. Figure 1.3 conceptually illustrates this idea from a Bayesian perspective. The prior distribution, shown at the top, represents the initial uncertainty about the unknown parameter before any measurements are taken. When design A is selected, the resulting posterior distribution becomes moderately narrower, indicating a partial reduction in uncertainty. In contrast, selecting a more informative design B yields a posterior distribution that is significantly narrower, corresponding to a greater reduction in uncertainty and thus a higher information gain.

To assess the quality of an experimental design, a *utility function* is introduced to quantitatively measure its effectiveness. Within the Bayesian framework, a natural and widely used choice of utility is the *Expected Information Gain* (EIG), which quantifies the anticipated reduction in uncertainty about the unknown parameters following data acquisition [53, 67, 64]. Conceptually, the EIG compares the entropy of the posterior distribution with that of the prior distribution, thereby measuring how much information is gained through the experiment. The design parameters that maximize this expectation are considered optimal. Equation 1.3 presents the corresponding formulation, where the joint distribution $\pi_{\text{joint}}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta})$ factorizes as $\pi_{\text{prior}}(\boldsymbol{x})\, \pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$.

$$\text{EIG}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\pi_{\text{joint}}(\cdot|\boldsymbol{\theta})} \left[\log \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta}) - \log \pi_{\text{prior}}(\boldsymbol{x})\right]. \tag{1.3}$$

In classical OED, so-called *alphabetic criteria* summarize the quality of a design in terms of the Fisher information matrix [43, 28, 56]. A well-known example is the *D-optimality criterion*, which seeks to maximize the determinant of the Fisher information matrix, thereby minimizing the volume of the confidence ellipsoid of the estimator. In the Bayesian setting, under the assumptions of a linear forward model with Gaussian noise (and typically a Gaussian prior), the D-optimality criterion can be shown to coincide with maximizing the expected information gain [14].

However, solving Equation 1.3 is generally highly challenging, as it involves a nested expectation (or equivalently a bi-level optimization problem) that is typically non-

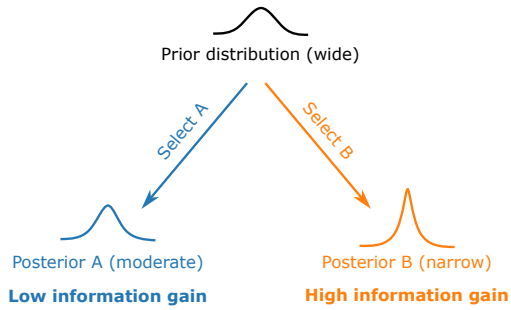


Figure 1.3: Conceptual illustration of information gain in Bayesian experimental design. A wide prior distribution represents initial uncertainty about the unknown parameter. Two different measurement designs, A and B, lead to distinct posterior distributions. Design B yields a narrower posterior—corresponding to reduced uncertainty and thus higher information gain.

convex and non-linear [66, 81]. In practice, the EIG is generally intractable. If both $\pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ and $\pi_{\text{prior}}(\boldsymbol{x})$ admit closed-form expressions, one can approximate the EIG using *nested Monte Carlo* sampling, though this approach is computationally expensive [64]. To improve efficiency, variational approaches have been proposed to approximate either the posterior $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta})$ or the marginal likelihood $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$, thereby yielding tractable lower or upper bounds on the EIG [32, 31].

In reality, however, neither $\pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ nor $\pi_{\text{prior}}(\boldsymbol{x})$ is typically available in closed form. A common simplification is to approximate the posterior distribution by a Gaussian using the *Laplace approximation*, which leads to a local quadratic approximation of the log-posterior around the MAP estimate. Under this approximation—often combined with a Gaussian prior—the Bayesian EIG reduces to a linear-model D-optimality criterion [39, 7, 26]. However, such approximations are inherently local and fragile, and they fail to capture the non-Gaussian structure that arises in many practical CT inverse problems.

1.2.3 Simultaneous vs. Sequential OED

In *simultaneous* (or *static*) OED, all design parameters are selected in a single step before any data are observed. In contrast, *sequential* (or *adaptive*) OED incorporates information obtained during the experiment, updating the design strategy as new measurements become available. In this case, the EIG is conditioned on the history of design choices and measurements [77, 64]. At the k^{th} step, given the history $\boldsymbol{h}_{k-1} =$

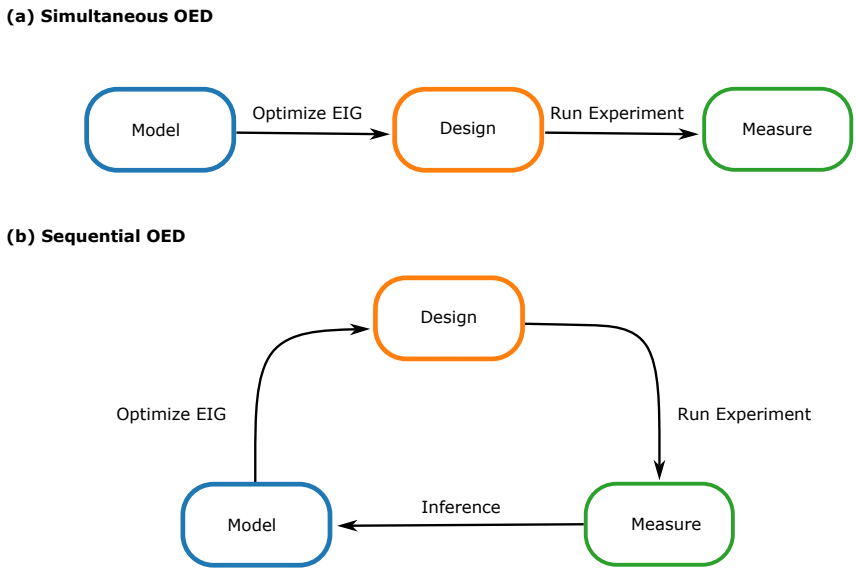


Figure 1.4: Comparison between simultaneous (static) and sequential (adaptive) OED. In simultaneous OED (a), all design parameters are selected in advance, and measurements are collected without adaptation. In sequential OED (b), design parameters are chosen step by step, with each choice conditioned on previously acquired data, enabling feedback-driven adaptation.

$\{(\theta_t, \boldsymbol{y}_t)\}_{t=1}^{k-1}$, the conditional EIG is defined as

$$\mathrm{EIG}(\theta_k \mid \boldsymbol{h}_{k-1}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_k)\sim\pi_{\mathrm{joint}}(\cdot\mid\theta_k,\boldsymbol{h}_{k-1})}\Big[\log \pi_{\mathrm{post}}(\boldsymbol{x} \mid \boldsymbol{y}_k, \theta_k, \boldsymbol{h}_{k-1}) - \log \pi_{\mathrm{prior}}(\boldsymbol{x} \mid \boldsymbol{h}_{k-1})\Big]. \tag{1.4}$$

where the joint distribution factorizes in the same way as in the simultaneous case, but using the updated prior:

$$\pi_{\mathrm{joint}}(\boldsymbol{x}, \boldsymbol{y}_k \mid \theta_k, \boldsymbol{h}_{k-1}) = \pi_{\mathrm{prior}}(\boldsymbol{x} \mid \boldsymbol{h}_{k-1})\, \pi_{\mathrm{data}}(\boldsymbol{y}_k \mid \boldsymbol{x}; \theta_k).$$

In most cases, the forward model π_{data} does not depend explicitly on the history $\boldsymbol{h}_{k-1}$ and can therefore be written as $\pi_{\mathrm{data}}(\boldsymbol{y}_k \mid \boldsymbol{x}; \theta_k)$. Sequential OED enables adaptivity by exploiting feedback from past measurements, but it also amplifies the computational challenges discussed previously, since sampling and approximations must be performed at every step of the experiment.

The optimization problems for simultaneous and sequential OED can be written as

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathrm{EIG}(\boldsymbol{\theta}), \tag{1.5}$$

$$\theta_k^* = \arg\max_{\theta_k} \mathrm{EIG}(\theta_k \mid \boldsymbol{h}_{k-1}). \tag{1.6}$$

As shown in Equations 1.5 and 1.6, optimal designs can be selected either in one shot or in a sequential manner. To make this optimization tractable, an *acquisition function* must be chosen. Under typical independence assumptions, the total information gain is a monotone submodular function of the selected design parameters. Consequently, *greedy search* methods can be used to approximate the optimal design: they are guaranteed to achieve at least a $(1 - 1/e) \approx 63\%$ fraction of the optimal EIG [59, 75, 50]. Alternatively, *Bayesian optimization* [34, 35] treats the EIG as a black-box objective, providing a flexible means to balance exploration and exploitation for efficient design selection in practice.

1.3 Challenges in Adaptive CT Acquisition

In inline inspection using X-ray CT, adaptive data acquisition is highly desirable because objects may vary in shape, size, and orientation as they move along the conveyor belt. Moreover, when the goal is defect detection, the acquisition strategy should adapt dynamically to the presence and characteristics of defects as they become visible in the reconstructed images. To accommodate such variability, feedback from previously acquired measurements should be incorporated into the scanning strategy. This naturally motivates the use of a sequential OED framework, which aligns well with the requirements of adaptive CT scanning.

While the theoretical framework of OED provides a principled foundation, several practical challenges limit its direct applicability to CT. First, as discussed in Section 1.2.2, neither the likelihood distribution $\pi_{\mathrm{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ nor the prior $\pi_{\mathrm{prior}}(\boldsymbol{x})$ is typically available in closed form for realistic CT problems. Second, CT reconstruction is inherently high-dimensional, making it computationally demanding to approximate models or perform sampling. In a sequential setting, these difficulties are compounded because inference must be performed at every step of the experiment.

Furthermore, inline inspection requires rapid decision-making: the acquisition function must be evaluated online to determine the next design parameter in real time. This poses difficulties for methods such as *greedy search* or *Bayesian optimization* [34, 35], which are often computationally expensive. Moreover, in many industrial

scenarios, the acquisition strategy must adapt to task-specific objectives, such as concentrating measurements on regions where defects or other abnormalities are likely to occur. Such objectives are difficult to encode directly into the information-gain framework, further complicating the design of adaptive CT acquisition strategies. These limitations motivate the development of a data-driven, model-free framework capable of learning efficient and task-aware acquisition policies.

1.4 Sequential OED with Reinforcement Learning

Reinforcement learning (RL) is a branch of machine learning that focuses on sequential decision-making. It has achieved remarkable success in diverse domains, including video games [57, 76, 86], board games such as AlphaGo [80], and large language models [60, 37]. The core idea of RL is to learn through interaction with an environment—by trial and error—how to choose actions that maximize cumulative rewards (utilities) over time. In this thesis, we apply the same principle to the sequential OED problem in CT, treating the scanning process analogously to playing a game by simulation. The agent learns to maximize the information gained through adaptive acquisition decisions. Once trained, the learned policy can be directly deployed to guide data acquisition.

1.4.1 Formulating Sequential OED as an RL Problem

RL offers a new perspective on the sequential OED problem by avoiding the reliance on closed-form probabilistic models required in traditional approaches. This allows RL-based methods to bypass the need for tractable priors, likelihoods, and posterior approximations required in classical sequential OED, enabling fast and adaptive acquisition strategies suitable for CT applications.

RL is grounded in the framework of a *Markov Decision Process* (MDP), which consists of an *action space*, *state space*, *transition function*, and *reward function* [84]. In practice, MDPs are often extended to *Partially Observable MDPs* (POMDPs), where the underlying parameters are not directly observable and must instead be inferred through belief states reconstructed from measurements. This formulation closely matches the setting of CT, in which the true object is only indirectly observed through noisy measurements. On the other hand, the *policy* in RL serves as an acquisition function, which directly maps the current reconstruction to the probabilities over the action space and can be parameterised. The goal of RL is to learn *a parameterized*

policy that maps the current state to the next action so as to maximize the expected cumulative reward. Through its probabilistic formulation, the policy naturally balances *exploration*—sampling novel or uncertain design parameters—and *exploitation*—selecting actions expected to yield high rewards. Accordingly, the optimization problem of sequential OED can be reformulated as learning a parameterized policy $\pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$, where $\boldsymbol{w}_a$ denotes the policy parameters. The *non-greedy* optimal policy π_{a}^*, with parameters $\boldsymbol{w}_a^*$, maximizes the expected cumulative reward obtained throughout the experiment.

Figure 1.5 illustrates the interaction loop between the reconstruction environment and the RL agent at each step. In the context of sequential OED, at the k^{th} step the *action space* corresponds to the possible values of the design parameter θ_k. After selecting θ_k, acquired projection sampled from the forward model is $\boldsymbol{y}_k \sim \pi_{\mathrm{data}}(\boldsymbol{y} \mid \bar{\boldsymbol{x}}; \theta_k)$. The reconstruction $\widehat{\boldsymbol{x}}_{k+1}$ is sampled from $\pi_{\mathrm{post}}(\boldsymbol{x} \mid \boldsymbol{y}_k, \theta_k, \boldsymbol{h}_{k-1})$ serves as the *belief state*. The *reward function* R corresponds to the EIG, which evaluates reconstruction accuracy. As the ground truth $\bar{\boldsymbol{x}}$ is often available from Computer-Aided Design (CAD) models or blueprints in industrial inspection, various metrics can be used to define R. For example, the reward can be the Peak Signal-to-Noise Ratio (PSNR) between the reconstruction and the ground truth for global reconstruction quality, or a contrast-based metric for task-specific objectives such as defect detection. This flexibility in reward design contrasts with the fixed EIG formulation of Bayesian OED.

By employing RL, we optimize *policy parameters* rather than the design parameters directly. We consider a finite-horizon return $\sum_{k=1}^{M} r_k$. Assuming the initial state is fixed, the objective function for RL is the expected cumulative reward from the initial state $\widehat{\boldsymbol{x}}_1$ under the trajectory distribution π_{chain}, which factorizes into the action policy and the data-driven transition:

$$J(\boldsymbol{w}_a) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_{\mathrm{chain}}} \left[\sum_{k=1}^{M} r_k \mid \widehat{\boldsymbol{x}}_1 \right], \tag{1.7}$$

where the trajectory distribution is

$$\pi_{\mathrm{chain}}(\boldsymbol{\tau}; \boldsymbol{w}_a) = \prod_{k=1}^{M} \pi_{\mathrm{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \cdot \pi_t(\widehat{\boldsymbol{x}}_{k+1} \mid \widehat{\boldsymbol{x}}_k, \theta_k), \tag{1.8}$$

π_t denotes the *transition function* in the standard RL formulation. In the CT context, it describes how the next reconstruction $\widehat{\boldsymbol{x}}_{k+1}$ is obtained after selecting an action

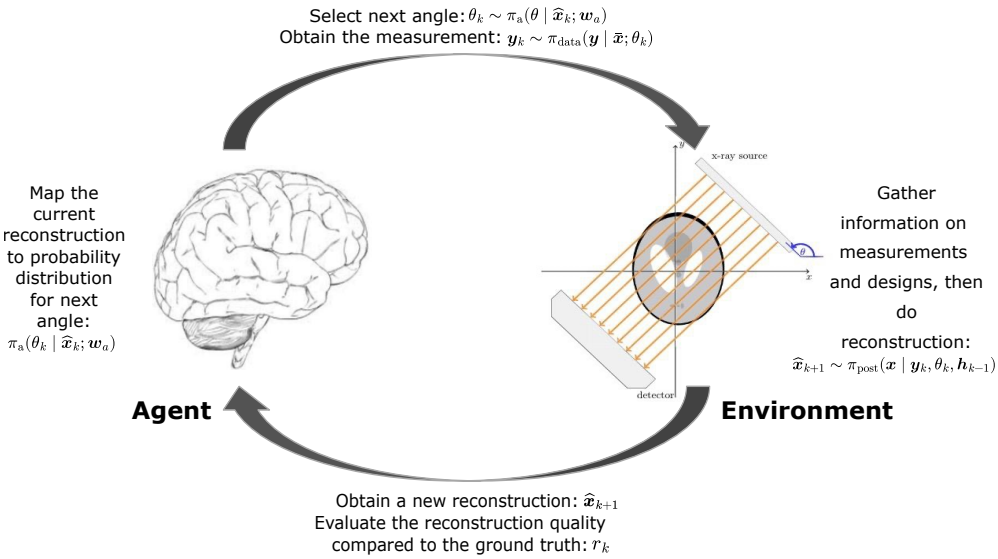


Figure 1.5: Acquisition–reconstruction loop at step k. **(1)** The current reconstruction (belief state) $\widehat{\boldsymbol{x}}_k$ is fed to the policy, which either terminates or selects the next angle θ_k. **(2)** If continuing, the corresponding measurement $\boldsymbol{y}_k$ is acquired. **(3)** The reconstruction is updated with the new data to obtain the refined estimate $\widehat{\boldsymbol{x}}_{k+1}$. **(4)** The reward r_k is computed from the improvement in a utility function, such as PSNR with respect to the ground truth $\bar{\boldsymbol{x}}$ (used only during training). Time flows from left to right as $k \to k{+}1$, with no feedback to earlier steps.

θ_k, reflecting the combined effect of acquiring a new measurement and updating the reconstruction. Analytically, this transition corresponds to applying the measurement likelihood (acquiring the projection at angle θ_k) followed by the reconstruction operator that updates the estimate using all available measurements. Because closed-form expressions for this transition are unavailable in realistic CT settings, we adopt a *model-free* approach and sample the transition by executing the forward projection at θ_k and performing one reconstruction update to obtain $\widehat{\boldsymbol{x}}_{k+1}$. A trajectory $\boldsymbol{\tau}$ of M steps, $\{\widehat{\boldsymbol{x}}_1, (\theta_1, \widehat{\boldsymbol{x}}_2, r_1), \ldots, (\theta_M, \widehat{\boldsymbol{x}}_{M+1}, r_M)\}$, is generated by iteratively selecting and evaluating actions according to the policy.

Two fundamental functions characterize the expected performance of a policy in RL. The *state-value function* $V^{\pi_a}(\widehat{\boldsymbol{x}}_k)$ denotes the expected cumulative reward obtained when starting from state $\widehat{\boldsymbol{x}}_k$ and following the policy π_a thereafter. When the initial state is fixed, $V^{\pi_a}(\widehat{\boldsymbol{x}}_1)$ coincides with the RL objective defined in Equation 1.7. The *state–action value function* $Q^{\pi_a}(\widehat{\boldsymbol{x}}_k, \theta_k)$ quantifies the expected cumulative reward obtained by taking action θ_k in state $\widehat{\boldsymbol{x}}_k$ and subsequently following the policy π_a.

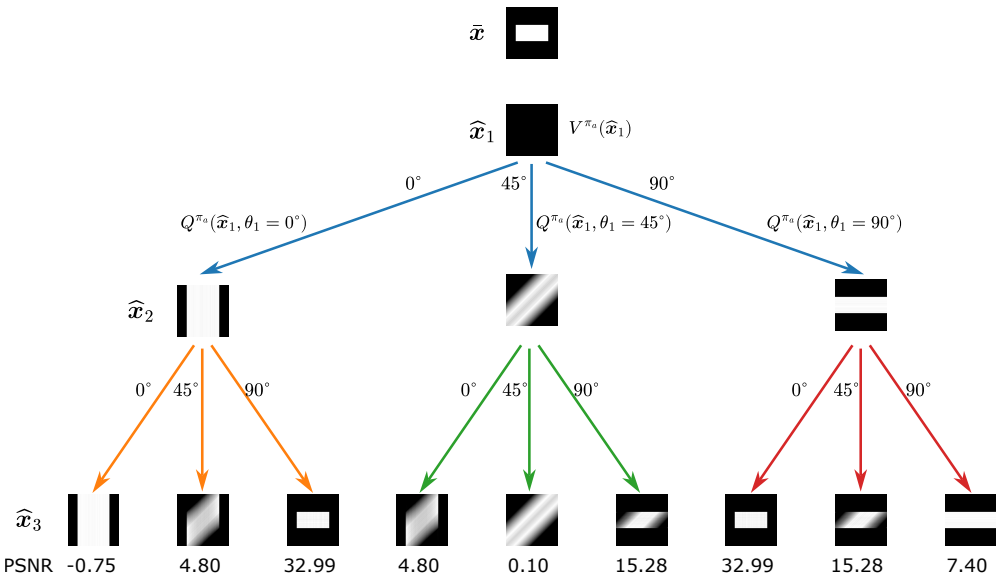


Figure 1.6: Illustration of the RL framework for projection angle selection in X-ray CT. The objective is to reconstruct the ground-truth image $\bar{\boldsymbol{x}}$ by selecting two projection angles from the candidate set $\{0°, 45°, 90°\}$. The figure shows the reconstruction results obtained for different pairs of selected angles, with reconstruction quality quantified by the PSNR. A uniform policy is applied, assigning equal probability to each action branch. Branches corresponding to different reconstruction paths are shown in distinct colors. The state value and action–state value under this uniform policy, corresponding to the initial state $\widehat{\boldsymbol{x}}_1$ (initialized as a zero matrix) and the state after the first angle selection, are highlighted.

Figure 1.6 illustrates the RL framework using a simple rectangular phantom as an example. The action space consists of three projection angle candidates, $\{0°, 45°, 90°\}$. Because the experiment is limited to two acquisition steps, the agent must select two angles from this candidate set. The figure shows the reconstructions obtained for all possible angle pairs, with reconstruction quality measured by the PSNR. In this example, the PSNR is used as the reward only at the final step, while intermediate steps receive zero reward. Consequently, the blue branches (which correspond to the first step) have reward 0, whereas the second-level branches (orange, green, and red) contain the PSNR values of the final reconstructions.

A uniform policy is applied, assigning equal probability $\frac{1}{3}$ to each action at every decision point. For simplicity, we additionally assume that the reconstruction resulting from selecting a single angle is deterministic. Under this policy, the action–state value $Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1)$ represents the expected cumulative reward obtained after taking action θ_1 in the initial state $\widehat{\boldsymbol{x}}_1$ (initialized as a zero image), averaged over all possible second-

step actions. For example, choosing $\theta_1 = 0°$ leads to the three orange branches, and

$$Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 0°) = \tfrac{1}{3}(-0.75 + 4.80 + 32.99) = 12.35.$$

Similarly, the action–state values for the other first-step actions are

$$Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 45°) = 6.73, \qquad Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 90°) = 18.56.$$

The state-value function aggregates these action–state values under the uniform policy:

$$V^{\pi_a}(\widehat{\boldsymbol{x}}_1) = \tfrac{1}{3}\big(Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 0°) + Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 45°) + Q^{\pi_a}(\widehat{\boldsymbol{x}}_1, \theta_1 = 90°)\big) = 12.55.$$

The state-value and state–action value for the second selection can be computed analogously. From these results, we observe that, under the uniform policy, selecting either 90° or 0° as the first projection angle yields higher expected returns. This indicates that the uniform policy is suboptimal and should be improved toward a non-uniform policy that assigns higher probability to selecting 0° and 90°, thereby increasing the value $V^{\pi_a}(\widehat{\boldsymbol{x}}_1)$.

To optimize the policy parameters, several policy-gradient methods may be used. A particularly effective approach is the *actor–critic* method, in which the policy (actor) is updated using estimates of the advantage provided by a learned value function (critic). The policy gradient can be approximated as

$$\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{k=1}^{T} \nabla_{\boldsymbol{w}_a} \log \pi_a(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \left(Q^{\pi_a}(\widehat{\boldsymbol{x}}_k, \theta_k) - V^{\pi_a}(\widehat{\boldsymbol{x}}_k) \right) \right], \quad (1.9)$$

where the critic provides an estimate of the advantage $A^{\pi_a}(\widehat{\boldsymbol{x}}_k, \theta_k) = Q^{\pi_a}(\widehat{\boldsymbol{x}}_k, \theta_k) - V^{\pi_a}(\widehat{\boldsymbol{x}}_k)$, which reduces variance and stabilizes the gradient updates. In practical implementations, both the state-value function and the state–action value function are learned using function approximators such as neural networks [84].

In the illustrative example shown in Figure 1.6, the horizon length is $T = 2$, corresponding to the requirement of selecting two projection angles. The parameter N denotes the number of sampled trajectories (i.e., complete angle-selection experiments) used to compute the Monte Carlo estimate of the policy gradient in Equation 1.9. Each trajectory produces a single realization of the sequence of rewards and the corresponding advantage values, and averaging these realizations over N trajectories provides an

empirical estimate of the gradient direction used to update the policy parameters.

Of course, this toy example contains only three actions and a short decision horizon, resulting in a very small decision tree. In realistic CT applications, however, the situation is far more complex: the objects being scanned are diverse, exhibiting variations in shape, position, scale, and orientation, and the state space consists of high-dimensional images. Moreover, the action space may be large, and the decision horizon can extend over many acquisition steps. Consequently, the underlying decision tree becomes exponentially large and cannot be enumerated. To address this complexity, neural networks are used to represent both the policy and the value functions, leading to the framework known as *deep reinforcement learning*, which enables scalable policy learning in high-dimensional CT reconstruction tasks.

In summary, reinforcement learning offers a principled and scalable framework for sequential experimental design in CT. In contrast to classical Bayesian OED approaches—which rely on explicit evaluation or approximation of the expected information gain—RL learns acquisition policies directly through interaction with the environment and can naturally incorporate task-specific objectives through the reward function. This black-box formulation is well suited to realistic CT systems, where accurate forward or inverse models and precise noise statistics may be unavailable or computationally expensive to evaluate. When such models are available or can be approximated reliably, RL can further benefit from a *model-based* formulation to improve sample efficiency. Moreover, the adaptive nature of RL policies enables real-time decision-making, making them particularly appealing for inline industrial inspection, where rapid feedback and efficient use of limited resources are essential. These considerations motivate the use of RL as the core methodological framework in this thesis.

1.5 Research Questions

This thesis investigates the use of RL for sequential OED in X-ray CT. The overarching goal is to develop and validate adaptive acquisition strategies that can be applied to inline CT inspection, where rapid and task-driven decision-making is essential. The research progresses from proof of concept to increasingly realistic and complex scenarios across the following chapters.

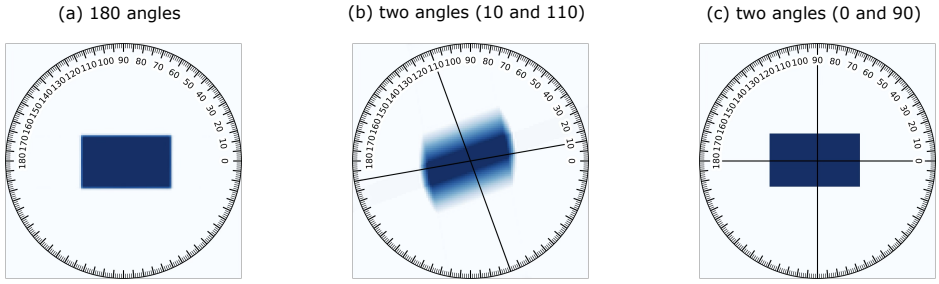


Figure 1.7: Reconstruction results using different angle selections. (a) Reconstruction from 180 uniformly distributed angles. (b) Reconstruction using two less informative angles (10° and 110°). (c) Reconstruction using two more informative angles (0° and 90°). Only the two informative angles achieve a reconstruction quality comparable to that of the full 180-angle case.

Research Question 1. *Can an RL agent learn to identify informative projection angles and achieve posterior-aware angle selection in X-ray CT?*

Figure 1.7 illustrates reconstruction results obtained using different angle selections. A parallel-beam geometry was employed, and reconstructions were computed using the Simultaneous Iterative Reconstruction Technique (SIRT) with a nonnegativity constraint over 150 iterations. For parallel-beam geometry, projection angles that are tangential to object edges tend to be more informative [102]. In Figure 1.7 (a), the reconstruction is obtained from 180 uniformly distributed angles, serving as the reference. Figures 1.7 (b) and (c) show reconstructions from only two projection angles: less informative ones (10° and 110°) and more informative ones (0° and 90°), respectively. Remarkably, using only two informative angles yields a reconstruction quality comparable to that obtained with the full set of 180 angles. This observation highlights the importance of identifying *informative* angles to address the challenges of efficient acquisition in industrial CT inspection.

In Chapter 2, we present a novel formulation of the projection-angle selection problem as a *Partially Observable Markov Decision Process* (POMDP), introduce an *Actor–Critic* reinforcement learning framework to address the sequential OED problem, and develop an adaptive acquisition strategy that can be efficiently applied to inline CT applications.

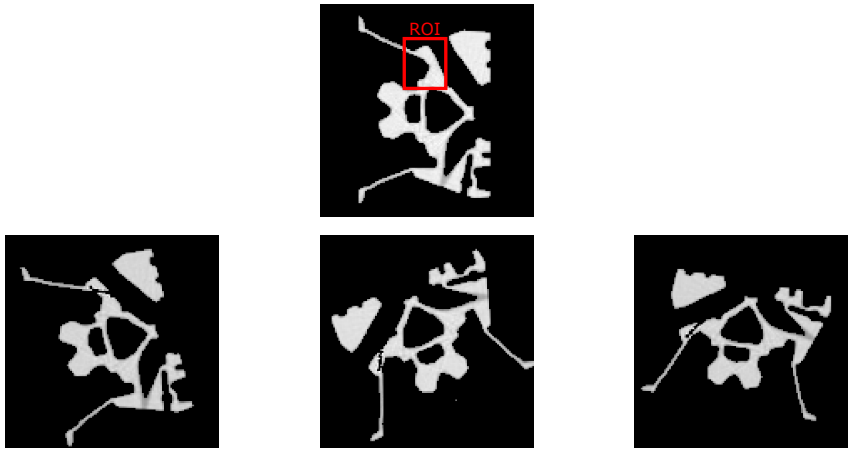


Figure 1.8: The top row depicts the selected ROI from the initial dataset, devoid of any defects. The bottom row presents three samples with artificially inserted crack defects within the ROI, each sample exhibiting a unique combination of rotation and scale variations to simulate defect diversity.

Research Question 2. *Can task-specific objectives, such as defect detection, be effectively integrated into the RL-based sequential design framework for inline CT?*

In Figure 1.8, assuming the defect-prone region is known, the reward function is adapted to focus on the corresponding Region of Interest (ROI), thereby emphasizing defect detectability. A segmentation-based evaluation is integrated into the acquisition pipeline to quantify how well the reconstructed image reveals the defect.

In Chapter 3, we extend this framework toward task-specific design by incorporating a downstream defect-detection objective directly into the RL formulation.

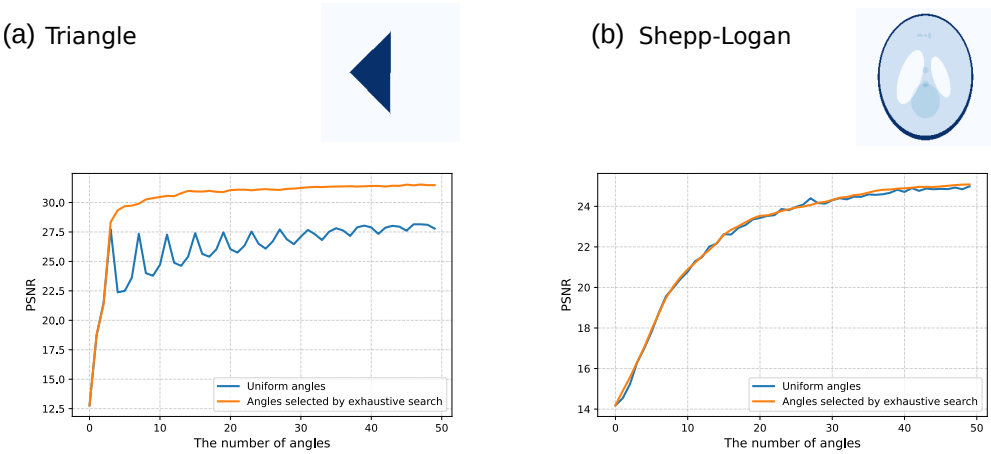


Figure 1.9: Noisy parallel-beam projections at 180 equally spaced angles; additive zero-mean Gaussian noise (5% level) is applied to the noise-free sinogram. PSNR as a function of the number of angles. The number of angles increases in increments of one. The orange curve represents angles selected via exhaustive search, while the blue curve corresponds to uniformly spaced angles. (a) Left: triangle phantom; (b) right: Shepp–Logan phantom.

Research Question 3. *Can an RL framework jointly learn an angle-selection policy and a terminal policy to adaptively determine the number of angles based on experimental cost while maintaining informative angle selection? Does a model trained on simulated data generalize effectively to similar experimental CT data?*

As shown in Figure 1.9, we plot reconstruction quality—measured by PSNR—versus the number of projection angles M (up to 50), with reconstructions obtained using SIRT under a nonnegativity constraint for 150 iterations. The orange curve corresponds to a selection that, at each step, exhaustively evaluates all remaining angles and chooses the most informative one; the blue curve uses uniformly spaced angles. For the triangle phantom (Figure 1.9 (a)), the selection consistently outperforms the uniform strategy, especially at small M. The uniform strategy occasionally picks informative directions, which explains the oscillations in its performance. For the Shepp–Logan phantom (Figure 1.9 (b)), which is approximately rotationally symmetric and lacks a preferred direction, the gap between the two strategies is small. Across both phantoms we observe diminishing returns: beyond a moderate number of angles, further increases in M yield only marginal PSNR gains. When acquisition is costly, such small gains may not justify additional projections, motivating an optimal stopping rule to balance reconstruction quality against experimental cost.

In Chapter 4, we address the *optimal stopping problem*. While previous work

assumed a fixed number of projection angles, here we aim to enable a flexible number of acquisition steps that balance reconstruction quality with experimental cost. To this end, we introduce a terminal policy and jointly optimize the policy gradients of the angle-selection and terminal policies. Additionally, we investigate the generalization capability of the learned policy by transferring it from simulated to experimental data.

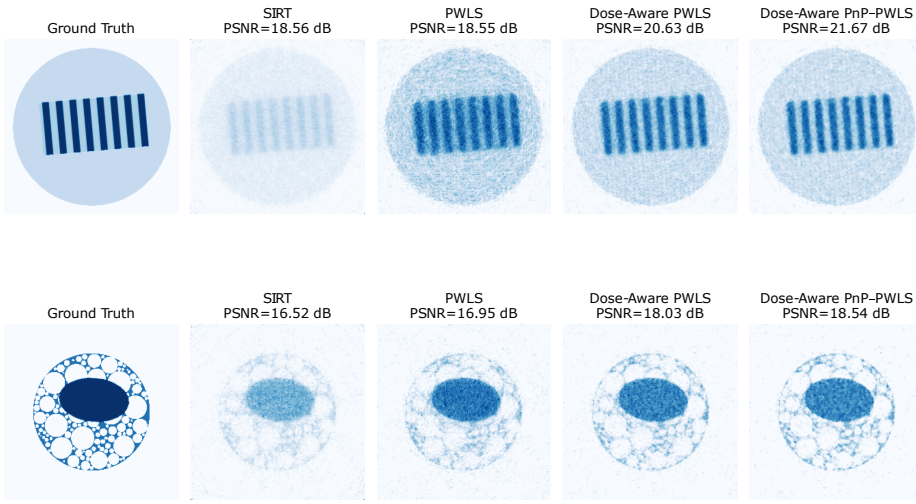


Figure 1.10: Reconstruction results for two phantoms with 60 projection angles using SIRT, conventional PWLS, and the proposed dose-aware methods. For each method, the iteration count was selected from 0 to 100 to yield its best reconstruction. SIRT reconstructions suffer from low contrast and artifacts, while PWLS provides modest improvements but remains noisy. The proposed dose-aware approaches achieve the highest PSNR and visibly improve contrast.

Research Question 4A. *How does a dose-aware reconstruction algorithm perform compared to conventional reconstruction under varying noise levels?*

Figure 1.10 compares four reconstruction methods: the Simultaneous Iterative Reconstruction Technique (SIRT), conventional Penalized Weighted Least Squares (PWLS), the proposed dose-aware PWLS formulation, and dose-aware PnP–PWLS incorporating a Plug-and-Play (PnP) TV denoiser. All reconstructions were performed under a non-negativity constraint for a fair comparison.

The visual reconstructions are consistent with the quantitative PSNR values shown in Figure 1.10. Across different noise levels, the dose-aware methods yield an improvement of approximately 1.5–3 dB over SIRT and conventional PWLS when considering their best results within the first 100 iterations. These results demonstrate that explicitly modeling angle-dependent dose allocation leads to substantially improved numerical performance and visibly cleaner, higher-contrast reconstructions—highlighting the importance of dose-aware modeling in CT.

Research Question 4B. *Can an RL agent jointly select projection angles and*

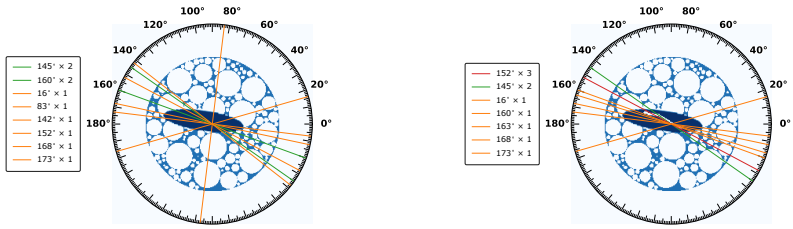


Figure 1.11: Angle-selection and dose-allocation strategies learned by the RL agent under a total dose budget of 2000 photons, shown separately for non-defective and defective phantoms. Each spoke indicates a parallel-beam projection direction at angle θ (degrees), measured counter-clockwise from the x-axis, where θ and $\theta + 180^\circ$ represent equivalent views. The color of each spoke encodes the number of times the corresponding angle is selected, which is proportional to the photon allocation at that angle. For instance, if one selection corresponds to 200 photons, then an orange spoke denotes an angle chosen once (200 photons), whereas a green spoke denotes an angle chosen twice (400 photons), and so forth.

allocate dose levels optimally to improve reconstruction quality and defect detectability under a limited total dose budget?

As illustrated in Figure 1.11, the RL policy adapts its acquisition strategy to the underlying object. For non-defective phantoms, the agent distributes projection angles and dose more uniformly, reflecting the absence of salient localized features. In contrast, for defective phantoms, the agent concentrates both angular sampling and photon allocation around views that are most informative for revealing the defect. This behavior indicates that the learned policy is able to jointly adjust angle selection and dose distribution in a task-aware manner under a fixed dose budget.

In Chapter 5, we further extend the framework by incorporating another key experimental parameter—*dose level*—which is critical for practical applications. The reconstruction algorithm is adapted to account for Poisson-distributed noise, and the RL agent is trained to allocate dose adaptively across different projection angles.

Chapter 2

Sequential Experimental Design for X-Ray CT Using Deep Reinforcement Learning

2.1 Introduction

X-ray Computed Tomography (CT) is a non-destructive method widely used to evaluate the quality of complex internal structures in industrial parts. However, there is a trade-off between high-quality reconstruction and scanning speed, as a time-consuming full 360-degree rotation is typically needed to obtain comprehensive information. Kazantsev [46] and Varga *et al.* [91] have pointed out that angles are not equally informative. Therefore, reducing the number of angles by extracting more informative data can help to improve the trade-off between reconstruction quality and scanning efficiency. This trade-off can be formulated as a bi-level optimization problem with respect to angle parameters [66]. The low-level optimization problem formulates the image reconstruction based on the chosen, limited projection data, while the high-level optimization problem finds angles that optimize the reconstruction quality.

This chapter is based on:

Tianyuan Wang, Felix Lucka, and Tristan van Leeuwen. "Sequential Experimental Design for X-Ray CT Using Deep Reinforcement Learning". In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 953–968. DOI: 10.1109/TCI.2024.3414273

2.1. Introduction

Bayesian Optimal Experimental Design (OED) is a mathematical framework that enables the acquisition of informative measurements while minimizing experimental costs [53, 67]. In Bayesian OED, the prior distribution represents the current belief about the underlying ground truth, while the posterior distribution refers to the updated belief after taking into account the new measurements obtained through the selected design. The difference between the prior and updated posterior reflects the change in uncertainty or equivalently the amount of information gained from the experiments. In *simultaneous* experimental design, we apply this procedure to select the optimal viewing angles in a single step, while in *sequential* experimental design, the goal is to select the viewing angles step-by-step, based on the projection data that has been collected so far [103]. It is this variant of the experimental design problem that we are interested in, as it can adapt the selected viewing angles to the object under investigation.

Two widely used methods for measuring the uncertainty reduction or information gain in Bayesian OED are D-optimality, and A-optimality [14]. D-optimality measures the information gain using the Kullback-Leibler divergence to compare the posterior and prior distributions, while A-optimality computes the expected error between the underlying ground truth and the reconstruction.

However, the high dimensionality, computational cost, and typically unknown or unobtainable prior distribution prevents the direct application of the aforementioned technique for sequential optimal design in real-time CT imaging.

Several methods have been proposed to address these issues. Implicit prior information has been the focus of some researchers. To this end, Batenburg *et al.* [9] and Dabravolski *et al.* [22] used a set of template images comprising Gaussian blobs to represent prior distribution samples and introduced an upper bound [8] to approximate the information gain, indicating the solution set's diameter. Gaussian distribution has been used as a tractable method for the prior distribution in [14]. Burger *et al.* sequentially selected the projection angle and the source-receiver pair's lateral position considering a specific region of interest and explored Bayesian A- and D-optimality to update the posterior in the covariance matrix and mean after each experimental step. Helin *et al.* [39] extended this work to non-Gaussian distributions and employed a Total Variation (TV) prior to enhance edges. In practice, a lagged diffusivity iteration generated a series of Gaussian approximations for the TV prior. Additionally, Barbano *et al.* [7] proposed a linearized deep image prior that incorporated information from the pilot measurements as a data-dependent prior. They then used a conjugate Gaussian-linear model to determine the next informative angles sequentially. How-

ever, these methods can be time-consuming and are not well-suited for fast in-line applications.

In an industrial context, the use of Computer-Aided Design (CAD) models is a common form of prior information. CAD models enable offline optimization by allowing angle acquisition using simulation tools. Fischer *et al.* [29] used a CAD model of the object to optimize task-specific trajectories based on the detectability index proposed by Stayman *et al.* [82]. The detectability index is computed using the modulation transfer function and noise power spectrum to evaluate its fitness with a user-defined frequency template. In addition to task-specific optimization, Herl *et al.* [42] considered data completeness optimization using a Tuy-based metric. Meanwhile, Bussy *et al.* [16] obtained a complete set of angles using either a simulation model or a CAD model and then used the discrete empirical interpolation method and related variants to sub-sample from the set of angles. Once a trajectory is optimized offline sequentially by a CAD model, it can be applied fast in the real application. Nonetheless, the alignment of the optimized trajectory outcome to the real-world coordinate system through proper registration is crucial before executing the real scan [10]. Hence, these methods lack genuine adaptability in in-line applications.

In terms of the methods discussed above, achieving adaptivity while maintaining a fast scan for in-line settings still presents a significant challenge. Additionally, informative angles are typically selected in a greedy manner after evaluating all available angle candidates. In the field of medical CT, Shen *et al.* [78] addressed this issue by training a deep reinforcement learning agent on a medical CT image dataset to personalize the scanning strategy sequentially. They utilized a gated recurrent unit as a policy network that maps all the previous measurements to a probability distribution over discrete angles and a radiation dose fraction. The next angle is chosen by sampling from this distribution. This way, around 60 are chosen sequentially.

We also leverage deep reinforcement learning to address the aforementioned challenges in our work but we focus on the application of industrial, in-line CT inspection instead of medical CT: We are considering very few scan angles (< 10), simple image features, but a potentially large inter-subject variation due to arbitrary placement and changing samples. For these reasons, we diverge from [78] by using the reconstruction space as the main state variable, avoiding problems caused by the increasing number of measurements. Due to this we use very different network architectures to parameterize the learned policy. By employing a deep reinforcement learning approach, we can train the policy to facilitate adaptive angle selection, offering a more efficient alternative to solving the high-dimensional, non-convex, bi-level optimization problem.

Figure (2.1) illustrates the proposed reinforcement learning approach for X-ray CT to solve this OED problem.

The contributions of this work include a novel formulation of the angle selection problem as a Partially Observable Markov Decision Process (POMDP), the use of the Actor-Critic approach from the field of reinforcement learning to address the OED problem, and the development of an adaptive approach that can be fast applied in in-line CT applications.

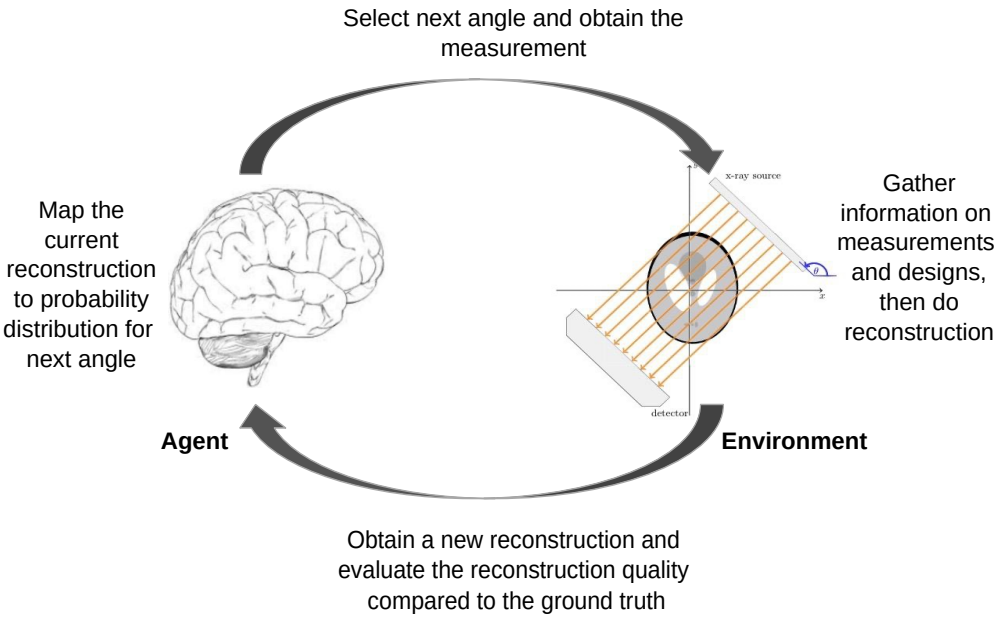


Figure 2.1: The interaction between the environment and the agent during policy training

The structure of this paper is as follows. In section 2.2, we present the background on CT reconstruction, Bayesian OED, and reinforcement learning. In section 2.3, we discuss the formulation of this experimental design as a POMDP and describe the computation of the policy gradient using the Actor-Critic approach. We provide a set of numerical experiments in section 2.4 to assess the performance of our proposed method. Finally, in section 2.5 and section 2.6, we discuss and summarize our findings.

2.2 Background

2.2.1 CT Reconstruction

In sparse-angle tomography, the challenge lies in accurately reconstructing an image from incomplete measurement data, where only a limited number of angles are acquired. This inverse problem is severely ill-posed, meaning that small errors in the measurements could result in a large reconstruction error, or that several reconstructions are consistent with the measurements [58]. The Filtered Back-Projection (FBP) algorithm, a traditional analytical reconstruction method, has limitations when used for sparse-angle tomography. It assumes that the measurements are acquired over the full angular range, resulting in inferior reconstructions when applied to limited data [38].

To address this challenge, it is necessary to incorporate prior information into the reconstruction algorithm to compensate for the limited data [17, 38]. Regularised algebraic reconstruction methods have been proposed to incorporate such prior information efficiently. When applied to limited data, these can result in more stable and accurate reconstructions.

Therefore, we represent the object that we would like to reconstruct as $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ where $n \in \mathbb{N}$ represents the number of pixels or voxels. A single noisy measurement $\boldsymbol{y}$ at angle θ is generated as

$$\boldsymbol{y}(\theta) = \boldsymbol{A}(\theta)\bar{\boldsymbol{x}} + \boldsymbol{\epsilon}, \tag{2.1}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{A}(\theta)$ is a discretization of the Radon transform along angle θ.

The reconstructed image from M measurements along angles $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$ is obtained via

$$\widehat{\boldsymbol{x}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{x}} \tfrac{1}{2} \sum_{k=1}^{M} \|\boldsymbol{A}(\theta_k)\boldsymbol{x} - \boldsymbol{y}(\theta_k)\|_2^2 + \alpha L(\boldsymbol{x}), \tag{2.2}$$

where $L(\boldsymbol{x})$ is a regularization term representing prior information for $\boldsymbol{x}$.

2.2.2 Bayesian OED

Bayesian OED is a statistical framework that optimizes the design of an experiment by trading off the information gain with the cost of an experiment.

In the context of X-ray CT experimental design, the utility function in Bayesian OED measures the reconstruction quality, where the true underlying ground truth $\bar{\boldsymbol{x}}$ is

estimated by $\widehat{\boldsymbol{x}}(\boldsymbol{\theta})$ from measurements $\boldsymbol{y} \in \mathcal{Y}$ obtained under experimental conditions specified by $\boldsymbol{\theta} \in \mathcal{D}$. The optimal design $\boldsymbol{\theta}^*$ maximizes the expectation of the utility function over the design space $\mathcal{D}$ with respect to the measured data $\boldsymbol{y}$ and the model parameter $\bar{\boldsymbol{x}}$.

Sequential OED is an approach that adjusts the design parameters as new data is acquired. This is achieved by treating the experiment as a sequential decision-making process, where the aim is to select the most informative design parameters based on the observed data to maximize the utility function. In the k_{th} step of an X-ray CT experiment, the process involves generating observed data using a data model $\pi_{\text{data}}(\boldsymbol{y}_k|\bar{\boldsymbol{x}};\theta_k)$ (as shown in Equation (2.1)), updating the posterior distribution of $\boldsymbol{x}$ given the observed data up to step k (denoted by $\pi_{\text{post}}(\boldsymbol{x}_k|\boldsymbol{y}_{1:k};\boldsymbol{\theta}_{1:k})$ in Equation (2.2)), obtaining the reconstruction for the underlying ground truth. Subsequently, the most informative angle θ_{k+1} is selected as the next design parameter to be used, which maximizes the utility function.

2.2.3 Reinforcement Learning

Reinforcement learning is a widely used approach for sequential decision-making, allowing agents to learn how to map the current state to actions that maximize the total reward for the entire process [84]. Since it considers the long-term effects of actions, reinforcement learning can realize non-greedy sequential decision-making. This approach is based on the Markov Decision Processes (MDPs) framework $\{\mathcal{S}, \mathcal{A}, \pi_t, R\}$, which consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition operator π_t representing the conditional probability distribution from the current state to the next state after selecting an action, and a reward function: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that provides feedback from the environment at each time step.

A policy π_{policy} in reinforcement learning is a mapping from the current state to a probability distribution of actions: $\pi_{\text{policy}}(a_k|\boldsymbol{s}_k)$.

In MDPs with a finite number of states, the process begins from an initial state $\boldsymbol{s}_1$ with a probability distribution $\pi_s(\boldsymbol{s}_1)$. The agent follows a policy that maps the initial state to the first action, leading the agent to transition to the next state and receive a reward from the environment. This process is repeated until a terminal state is reached, generating a trajectory or an episode $\boldsymbol{\tau} = (\boldsymbol{s}_1, a_1, r_1, ..., \boldsymbol{s}_M, a_M, r_M)$ of M steps.

A Partially Observable Markov Decision Process (POMDP) is an extension of a MDP. In many practical scenarios, an agent may not have full visibility or knowledge

about the environment's state. POMDPs come into play in such situations, allowing an agent to make decisions based on limited or partial observations. A POMDP can be defined as a tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{O}, \pi_t, \pi_e, R\}$, where two additional components are included in addition to the ones in the standard MDP formulation: a finite observation set $\mathcal{O}$ and an observation function π_e that defines the conditional probability distribution over the observation in the underlying state after executing an action. Since the agent has limited knowledge about the underlying state in POMDPs, the policy must either map historical observations to the next action or extract information from historical observations in the form of a belief state.

Reinforcement learning aims to find the optimal policy with parameters $\boldsymbol{w}$, denoted as $\pi^*_{\text{policy}}(.; \boldsymbol{w})$, that generates the trajectory or episode $\boldsymbol{\tau}$ to maximize the expected total reward. The objective function for reinforcement learning can be expressed as follows:

$$\begin{aligned} &J(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_{\text{chain}}} \sum_{k=1}^{M} \gamma^{k-1} r_k, \\ &\text{where } \pi_{\text{chain}} = \pi_s(\boldsymbol{s}_1) \prod_{k=1}^{M} \pi_{\text{policy}}(a_k | \boldsymbol{s}_k; \boldsymbol{w}) \pi_t(\boldsymbol{s}_{k+1} | \boldsymbol{s}_k, a_k). \end{aligned} \tag{2.3}$$

The objective function measures the expected total reward with a discount factor $\gamma \in (0, 1]$ to account for future uncertainty, and π_{chain} represents the trajectory generation process by the policy.

The total rewards for one trajectory are obtained after the agent completes an episode. The expectation over all trajectories can be estimated by sampling many trajectories. To enhance the process's efficiency, some reinforcement learning approaches utilize value functions that evaluate the expected future benefits from the k_{th} step following the policy. The state-value function $V(\boldsymbol{s}_k)$ quantifies the expected cumulative reward from state $\boldsymbol{s}_k$, taking into account all possible trajectories following the current policy that start from this state.

2.3 Methods

2.3.1 Sequential OED as a POMDP

We take the reconstruction as a belief state rather than considering measurements as the state, as done in [78]. To formulate the problem, we adopt a Bayesian OED framework and model it as a POMDP. The POMDP formulation for the X-ray CT

experiment is defined as follows:

- Observation space $\mathcal{O}$: The observation space is defined as the set of measurements generated by the data model expressed in Equation (2.1).
- State space $\mathcal{S}$: The ground truth $\bar{\boldsymbol{x}}$ represents the underlying state. The current reconstruction (belief state) of the underlying state, denoted by $\widehat{\boldsymbol{x}}(\boldsymbol{\theta}_{1:k})$, is obtained using the SIRT algorithm with box constraints [27] as specified in Equation (2.2). For ease of notation, we use $\widehat{\boldsymbol{x}}_k$ to represent the reconstruction at the k_{th} step. In addition, we maintain a vector $\boldsymbol{b}_k$ to keep track of the angles that have been selected before the k_{th} experiment to prevent repeating the same angles.
- Action space $\mathcal{A}$: The action space is a discrete design space consisting of 180 integer angles from the range $[0°, 180°)$.
- Transition function π_t and observation function π_e: The transition function π_t is deterministic, as the underlying state remains unchanged. On the other hand, the data model π_e given by Equation (2.1) serves as the observation function, from which we only consider measurement samples.
- Reward function R: The reward function is defined based on the PSNR value between the reconstruction obtained after selecting the angle θ_k and its ground truth. Two reward settings are considered, both of which correspond to A-optimality in Bayesian OED.:
  - End-to-end setting: The reward is given as follows:

$$R(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) = \begin{cases} PSNR(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) & \text{if } k = M \\ 0 & \text{otherwise} \end{cases}$$

    If the fixed number of angles M is reached, the episode terminates, and the final PSNR value is given. Otherwise, the agent receives a reward of 0.
  - Incremental setting: The reward is given as follows:

$$R(\widehat{\boldsymbol{x}}_{k+1}, \widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}}) = PSNR(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) - PSNR(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}})$$

    The reward represents the improvement in the current reconstruction quality compared to the previous step.

2.3.2 Actor-Critic method for policy optimization

The Actor-Critic method is a novel category in the field of reinforcement learning for computing the policy gradient on the objective function described in Equation (2.3). The proposed approach leverages the concept of value functions and utilizes a state-value function to obtain the expected future rewards at the current state, thereby expediting the learning process. Additionally, this method parameterizes the value function. The Temporal-Difference (TD) error [84] is employed in this approach, which calculates the discrepancy between the estimated value function for the current state and the sum of the current reward and the discounted estimated value function for the next state. This enables the state-value function to be updated through bootstrapping and provides a direction for policy gradient.

At the beginning of each episode, a zero matrix and a zero vector are used as the initial state and action vector, respectively. The complete algorithm is presented in Algorithm (1).

2.3.3 Network architecture

The proposed method requires the agent to extract relevant features from high-dimensional images to increase learning efficiency, which is accomplished using a deep neural encoder network. The architecture of the encoder network and the Actor-Critic network is shown in Figure (2.2), with the input image being of dimension 128×128. The neural network's connection weights represent the policy parameters $\boldsymbol{w}_1$ and the state-value function parameters $\boldsymbol{w}_2$.

The proposed model adopts a shared encoder network between the actor and critic networks. This shared encoder network comprises three convolutional neural networks (CNNs) each with padding and group normalization, followed by a leaky Rectified Linear Unit (ReLU) activation and a max pooling operation for down-sampling. The shared encoder network consists of a total of 13,320 parameters. Furthermore, the following actor and critic networks are separate and have 170,820 and 900,601 parameters, respectively.

Figure (2.2) outlines the process by which the network operates in the context of the Actor-Critic method. The encoder network takes the reconstruction $\widehat{\boldsymbol{x}}_k$ as input and produces a feature vector in the bottleneck layer, which is flattened into a 1D vector and concatenated with the 1D action vector $\boldsymbol{b}_k$. The resulting information is then fed into the following actor and critic networks.

The actor network uses a Soft-max policy to map the information to a probability

Algorithm 1 Actor-Critic.

Initialize the policy parameters $\boldsymbol{w}_1$ and the value function parameters $\boldsymbol{w}_2$ randomly. Set step sizes $\alpha^{\boldsymbol{w}_1} > 0$ and $\alpha^{\boldsymbol{w}_2} > 0$

for each episode do: **do**
    Get a phantom sample $\bar{\boldsymbol{x}}$ then initialize $\widehat{\boldsymbol{x}}_1 = \mathbf{0}$ and
    $\boldsymbol{b}_1 = \mathbf{0}$ (first state of this episode).
    for $k = 1, ..., M$:
        Select the angle based on the soft-max policy,
        which maps the inputs to a probability distribution
        that sums to 1: $\theta_k \sim \pi_{\text{policy}}(\cdot|\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_1)$
        Get new measurements $\mathbf{y}_k$ from Equation (2.1)

        Reconstruct new image $\widehat{\boldsymbol{x}}_{k+1}$ from Equation (2.2)
        and get a new vector $\boldsymbol{b}_{k+1}$

        Get reward r_k using end-to-end setting
        $R(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}})$ or incremental setting $R(\widehat{\boldsymbol{x}}_{k+1}, \widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}})$
        Estimate the state-values $\widehat{V}(\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_2)$ and
        $\widehat{V}(\widehat{\boldsymbol{x}}_{k+1}, \boldsymbol{b}_{k+1}; \boldsymbol{w}_2)$ using a neural network
        Compute TD error:
        $\delta_k \leftarrow r_k + \gamma\widehat{V}(\widehat{\boldsymbol{x}}_{k+1}, \boldsymbol{b}_{k+1}; \boldsymbol{w}_2) - \widehat{V}(\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_2)$
        (If $k = M$, $\widehat{V}(\widehat{\boldsymbol{x}}_{k+1}, \boldsymbol{b}_{k+1}; \boldsymbol{w}_2) = 0$)
        Update policy function parameters $\boldsymbol{w}_1$:
        $\boldsymbol{w}_1 \leftarrow \boldsymbol{w}_1 + \alpha^{\boldsymbol{w}_1}\nabla_{\boldsymbol{w}_1} \log \pi_{\text{policy}}(\theta_k|\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_1)\delta_k$
        Update value function parameters $\boldsymbol{w}_2$:
        $\boldsymbol{w}_2 \leftarrow \boldsymbol{w}_2 + \alpha^{\boldsymbol{w}_2}\nabla_{\boldsymbol{w}_2}\widehat{V}(\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_2)\delta_k$
    end for
end for

distribution over all possible angle candidates in the action space, while the critic network estimates the state-value function $\widehat{V}(\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_2)$. Based on the probability distribution generated by the actor network, the agent selects the next angle θ_k and subsequently collects measurements to obtain a new reconstruction $\widehat{\boldsymbol{x}}_{k+1}$. The action vector is updated as $\boldsymbol{b}_{k+1}$ accordingly.

To compute the policy gradient and update the parameters in the value function using TD error in Algorithm (1), the new reconstruction $\widehat{\boldsymbol{x}}_{k+1}$ and the new action vector $\boldsymbol{b}_{k+1}$ are fed into the network again. This is done to calculate a new state-value function $\widehat{V}(\widehat{\boldsymbol{x}}_{k+1}, \boldsymbol{b}_{k+1}; \boldsymbol{w}_2)$. Once an angle is selected, both the policy parameters $\boldsymbol{w}_1$ and the value function parameters $\boldsymbol{w}_2$ are updated once.

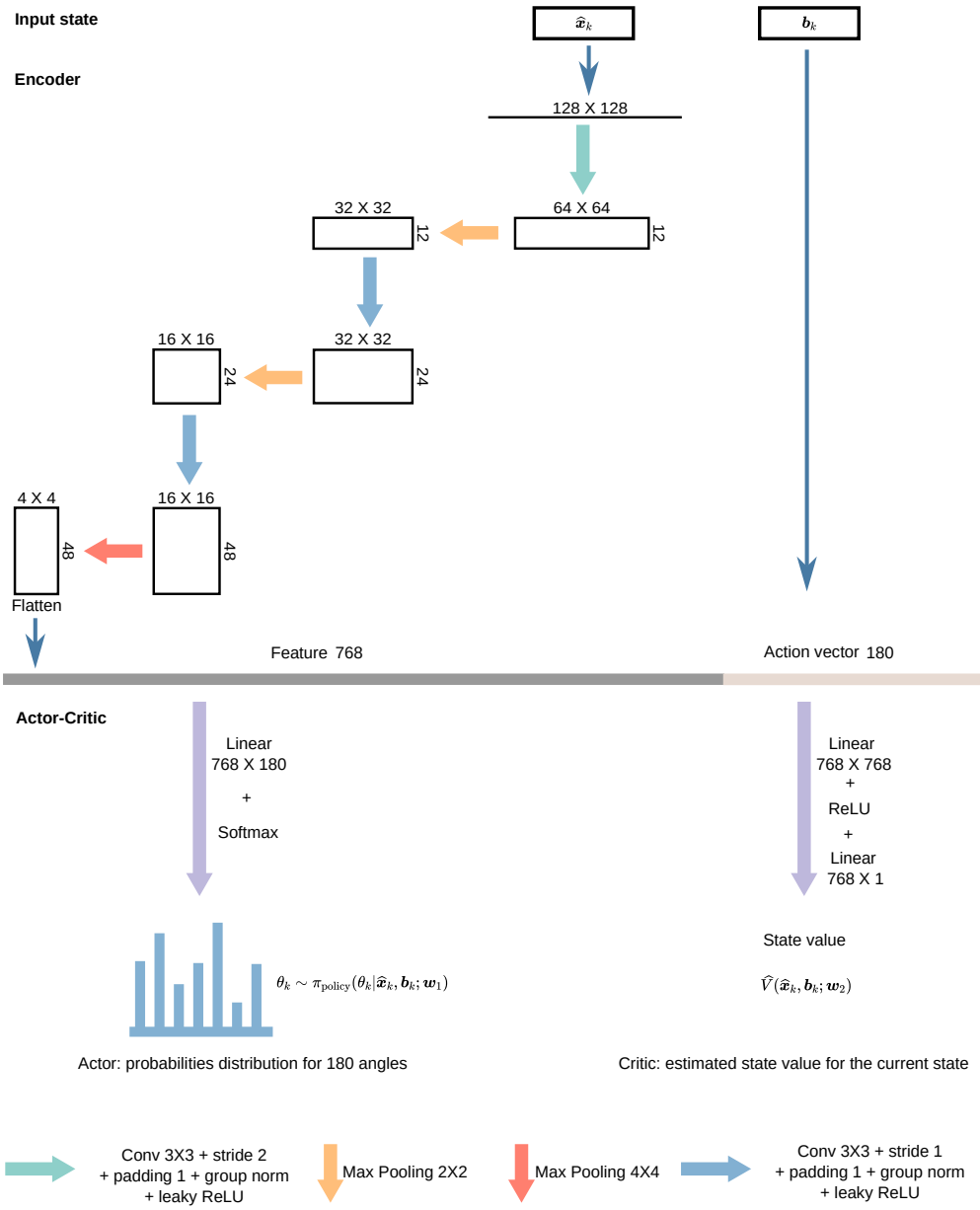


Figure 2.2: The combined network architecture first consists of an image encoder branch that processes the current reconstruction. Then, the code is concatenated with the previous action vector and fed into a network parameterizing the policy (Actor, left bottom) and a network estimating the state value (Critic, right bottom).

2.4 Numerical Experiments

We examine in some numerical experiments whether the learned policies are really able to sequentially adapt the scan angles to the object (a-posteriori adaptation). For this, we use various simple numerical phantoms—where informative angles are well-established—as well as more realistic phantoms. Throughout our experiments, we focus on parallel-beam geometry and 2D tomography using synthetic data. The code and synthetic data are available on Github [1].

2.4.1 Datasets

In our numerical experiments, we consider several shapes depicted in Figure (2.3). All phantoms have a size of 128×128. The phantoms in the datasets from d1) to d5) are binary images. To facilitate algorithm validation with more sophisticated scenarios, datasets d7) and d8) feature advanced phantoms composed of multiple materials. To assess the adaptability of the agent to dynamic environments, each dataset from d1) to d8) includes phantoms with different rotations, causing a shift in their informative angles. These rotations are represented by 36 equally spaced angles ranging from 0° to 179°. Additionally, the phantoms in each dataset from d1) to d8) exhibit various scaling and shifts (parameters are shown in Appendix A.1). Nonetheless, these modifications do not alter the informative angles, thereby preserving the consistency of informative angles across the scaled and shifted phantoms. By including these scaling and shifts, we aim to ensure the agent's ability to recognize the same object despite its size and location variations. Using these datasets allows us to transparently validate the algorithm.

d1) Circles: The first dataset consists of circles with varying locations and radii. Due to its uniform curvature, a circle does not have a relatively higher concentration of informative angles. To obtain an accurate reconstruction, angles must be equidistantly distributed.

d2) Ellipses: Unlike circles, ellipses have a major axis and a minor axis. The major axis serves as a preferential direction, making angles around it more informative, as shown in references [9] and [92].

d3) Triangles: Triangles, characterized by one angle of 90° and two angles of 45°, possess three preferential directions, causing the informative angles to be tangential to their edges [102].

[1] `https://github.com/tianyuan1wang/SeqAngleRL`

d4) Mixed phantoms: This dataset consists of a mixture of phantoms, including triangles from d3), regular pentagons, and regular hexagons, each of which has its own preferential directions.

d5) and d6) Combined phantoms: The dataset d5) comprises samples created by combining elements from two distinct phantom types found in categories d1), d2), and d4). In contrast, dataset d6) consists of two distinct samples from dataset d4), which does not include round shapes. These two datasets integrate features from a variety of shapes. Through this approach, the dataset transforms simple phantoms into more complex entities, facilitating a broader exploration of shape and structure.

d7) and d8) Complex phantoms: The datasets under consideration introduce variations to the classic Shepp-Logan phantoms. Specifically, we have redesigned the Shepp-Logan phantoms by excluding the outer ellipse and modifying the two inner ellipses, ranging from the lowest to the highest intensity values. In contrast to previous datasets, the current ones integrate both static and dynamic elements. In the d7) setup, two high-intensity ellipses experience changes in both scale and rotation across 36 uniformly distributed angles from 0° to 179°. Following such modifications, these redefined ellipses markedly impact the reconstruction error, consequently influencing the adaptive capabilities of the trained policy. On the other hand, in the d8) setup, only the lower of these ellipses exhibits dynamic characteristics.

2.4.2 Implementation

For all of our experiments, the sequential experimental process for each dataset in Figure (2.3) follows Algorithm (1). To generate the measurement data, we utilize the ASTRA Toolbox [90, 89], considering a projection size of 1.5×128. This specific size choice ensures that each angle covers all pixels. The reconstruction is performed using the SIRT algorithm with box constraints [0,1] for 150 iterations.

The encoder and Actor-Critic neural network architectures are illustrated in Figure (2.2). During training, we set the discount factor γ to 0.99 and assign weights of 1.0 and 0.5 to the actor loss: $-\log \pi_{\text{policy}}(\theta_k|\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_1)\delta_k$ and critic loss: $(\delta_k)^2$, respectively. To encourage exploration during training, we incorporate an entropy loss: $\sum_{\theta} \pi_{\text{policy}}(\theta_k|\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_1) \log \pi_{\text{policy}}(\theta_k|\widehat{\boldsymbol{x}}_k, \boldsymbol{b}_k; \boldsymbol{w}_1)$, with a weight of 0.01. The aforementioned parameter settings have been empirically determined to achieve an optimal balance between policy optimization, accurate value estimation, and robust exploration throughout the training regimen. For optimization of the parameters, we employ the Adam optimizer [47] with a learning rate of 10^{-4} and weight decay of 10^{-5}.

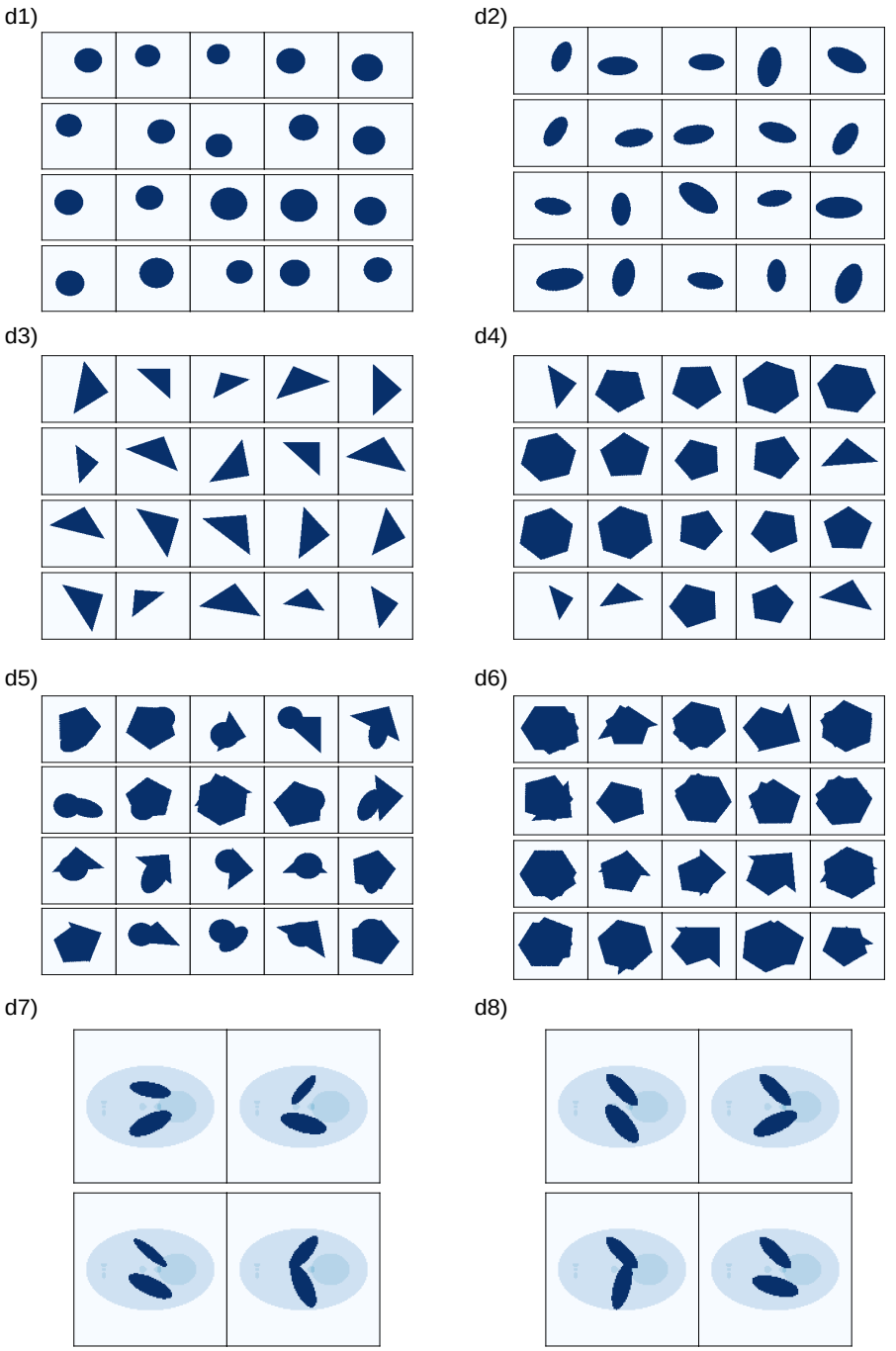


Figure 2.3: Datasets utilized in subsequent experiments: Each dataset comprises samples that exhibit a variety of rotations and scales.

In addition, we consider two reward functions: incremental and end-to-end settings. An equidistant policy is introduced as a benchmark to compare the performance of the Actor-Critic policy with un-informed and non-adaptive angle selection method. Significantly, Experiments 1, 2, 3, 4, and 6 employ distinct policies, each specifically tailored to their respective training datasets. Conversely, Experiment 5 adopts either the policy developed in Experiment 4 or a newly trained policy encompassing multiple datasets, to thoroughly evaluate and validate its ability to generalize.

2.4.3 Experiment 1 - Uniform informative angles

In the first experiment, we aim to train an agent and evaluate its performance on the circles (dataset d1), which have a uniform distribution of informative angles. It is known that the equidistant benchmark is the optimal policy for this dataset. Our objective is to investigate whether the Actor-Critic policy approaches the equidistant policy in performance. To facilitate this investigation, the experiment is conducted with 3,000 training phantoms, and it requires an extensive training duration of 100,000 episodes.

As depicted in Figure (2.4), the equidistant policy exhibits enhanced performance for the circular phantoms compared to the Actor-Critic policies with diverse reward configurations. Furthermore, we observe that the performance of the Actor-Critic policy with end-to-end reward surpasses that of the policy with incremental reward as the number of angles increases.

Figure (2.5) presents two samples considering three and seven angles obtained from the Actor-Critic policy with the end-to-end reward setting. This result demonstrates that the Actor-Critic agent tends to distribute the selected angles evenly, although the number of angles is different.

2.4.4 Experiment 2 - Non-uniform informative angles

In contrast to circles, informative angles in ellipses (dataset d2) are found to be concentrated around its major axis. We train an agent on a dataset comprising 3,000 ellipse phantoms. The model achieves convergence after undergoing 150,000 training episodes.

The training outcomes of the ellipse phantoms over the final 2,000 episodes are shown in Figure (2.6), which indicates that the Actor-Critic policies exhibit superior performance. As the number of angles increases, the results for the three policies get closer. This is because a sufficient number of angles around the major axis have already

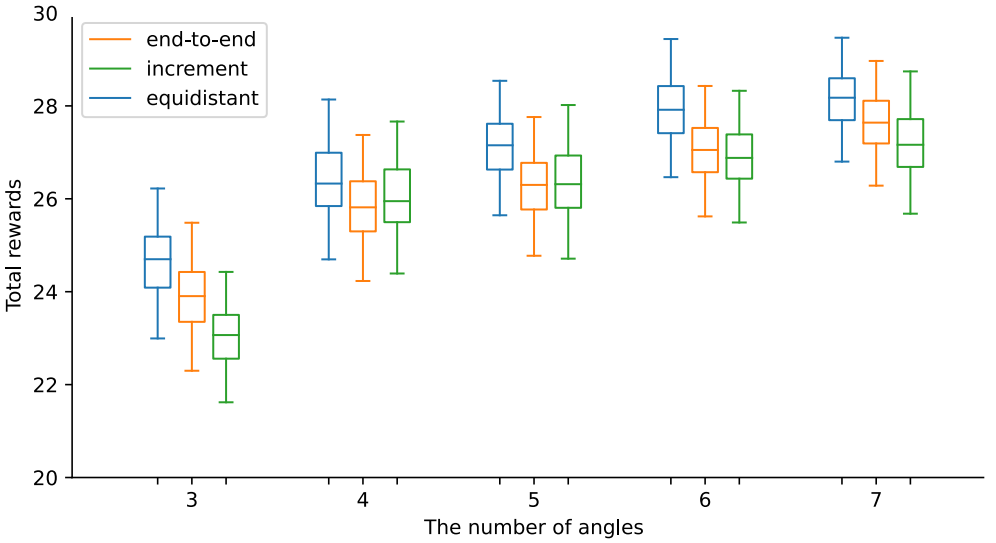


Figure 2.4: Comparison of policies considering different numbers of angles for the circles dataset: the results demonstrate the training outcomes over the last 2,000 episodes. The box represents the interquartile range in these plots, spanning from the first to the third quartile of the data distribution. The median value is displayed as a line within the box. The whiskers extend from the box to illustrate the range of the data distribution beyond the interquartile range.

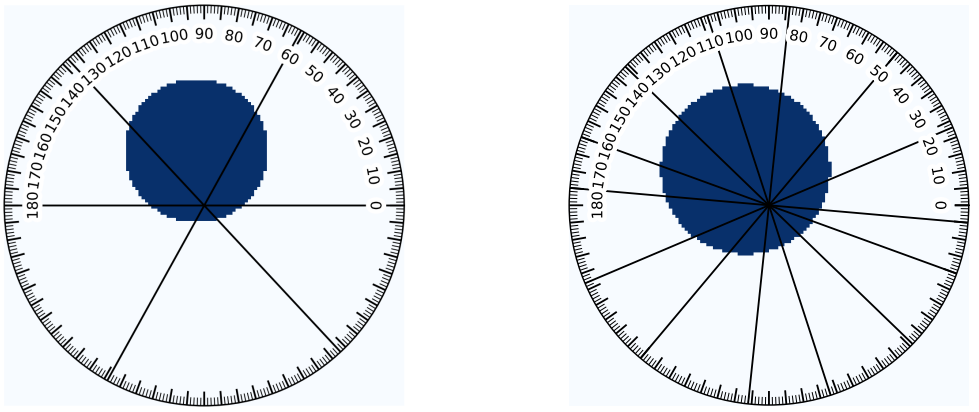


Figure 2.5: Results of the end-to-end reward setting for two circle phantoms considering three and seven angles.

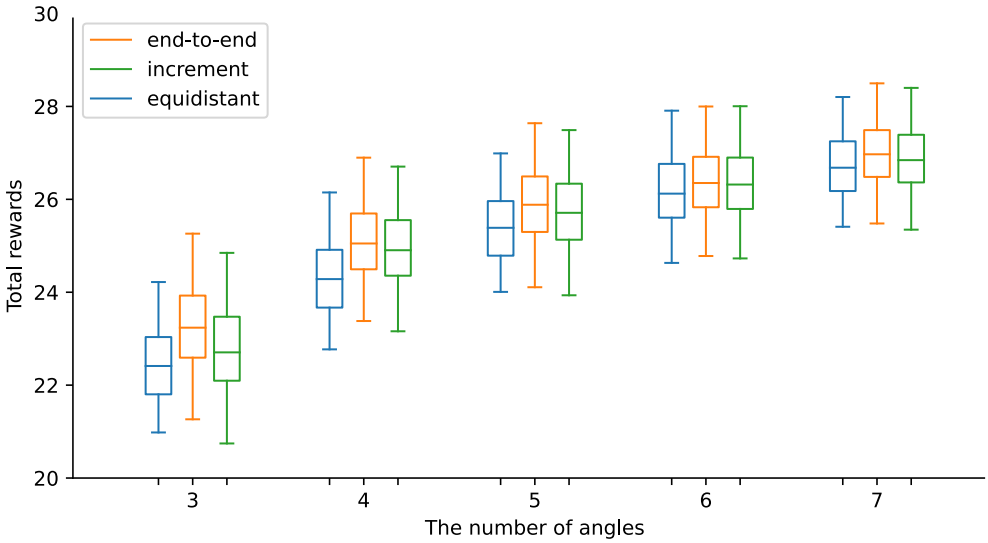


Figure 2.6: Comparison of policies considering different numbers of angles for the ellipses dataset: the results demonstrate the training outcomes over the last 2,000 episodes. The box represents the interquartile range in these plots, spanning from the first to the third quartile of the data distribution. The median value is displayed as a line within the box. The whiskers extend from the box to illustrate the range of the data distribution beyond the interquartile range.

Table 2.1: Performance comparison of policies on unseen rotations test for ELLIPSES regarding the PSNR values

Policies	3	4	5	6	7
Learned adaptive policy (end-to-end)	**23.16 ± 1.02**	**25.10 ± 0.72**	**25.73 ± 1.20**	26.31 ± 0.82	**26.87 ± 1.06**
Learned adaptive policy (increment)	22.78 ± 0.92	24.90 ± 0.73	25.67 ± 0.77	**26.33 ± 0.79**	26.86 ± 0.73
Equidistant policy	22.40 ± 0.74	24.27 ± 0.73	25.35 ± 0.69	26.16 ± 0.64	26.73 ± 0.62

been obtained, even for the equidistant policy, to achieve a high-quality reconstruction. Notably, the Actor-Critic policy with the end-to-end reward setting achieves the best performance. Figure (2.7) presents the results for two ellipse phantoms, demonstrating that the agent can discern the rotation of the ellipse and concentrate the distribution of the angles around the informative area. As the number of angles increases, the agent increases the number of angles around the major axis.

The unseen test set comprises 300 phantoms. Table (2.1) reports the outcomes regarding the unseen rotations for the ellipses dataset with three to seven angles. Consistent with the training results, the Actor-Critic policies demonstrate superior performance compared to the benchmark, with the policies becoming progressively closer as the number of angles increases. The end-to-end reward setting still shows the best average performance, though it has a higher variance.

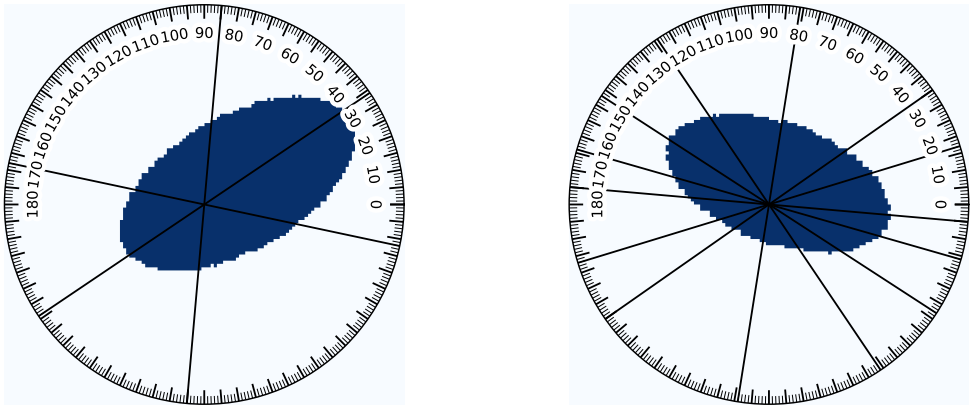


Figure 2.7: Results of the end-to-end reward setting for two ellipse phantoms considering three and seven angles.

2.4.5 Experiment 3 - Explicit informative angles

The third experiment focuses on evaluating the ability of the Actor-Critic agent to identify explicit informative angles for phantoms with sharp edges, namely triangles

(dataset d3). These phantoms have well-defined informative angles that are tangential to their edges, and thus, it is of interest to investigate if the agent can successfully locate these angles. The results of this experiment will provide insight into the performance of the Actor-Critic agent in detecting preferential directions for phantoms with sharp edges.

This study employs a dataset comprising 3,000 triangle phantoms for training the agent. A total of 150,000 episodes are required to reach convergence in the training process. A fixed number of five angles is employed. As shown in Figure (2.8), both reward settings for the Actor-Critic agent outperform the equidistant policy. Specifically, training using the incremental reward setting demonstrates faster convergence, whereas the end-to-end reward setting yields the best performance.

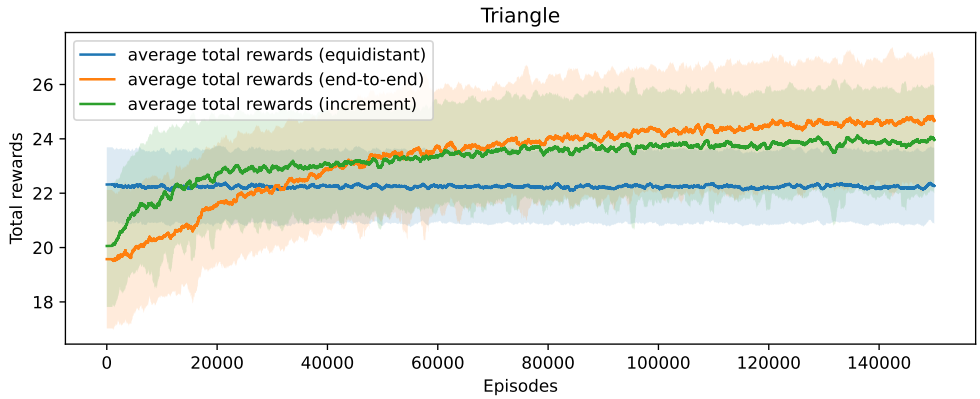


Figure 2.8: This figure compares the performance of Actor-Critic policies, trained on triangles phantoms, with that of an equidistant policy. It displays the training outcomes, with curves showing the mean values and shaded color bands representing the variances.

The training results demonstrate that the Actor-Critic agent tends to select the first two angles as fixed angles, with particular emphasis on the first angle, while the second angle exhibits some uncertainty. Subsequently, the agent would select three informative angles to optimize the reconstruction process. This behavior is consistent with the fact that the initial state is set as a zero matrix and a zero vector with no prior information, and the agent, therefore, prioritizes gathering information by fixing the first angle or first two angles before personalizing the strategies based on the different phantoms encountered.

Figure (2.9) presents two samples of the agent's performance in an end-to-end reward setting, in which the agent selects the initial two angles of 44° and 153°.

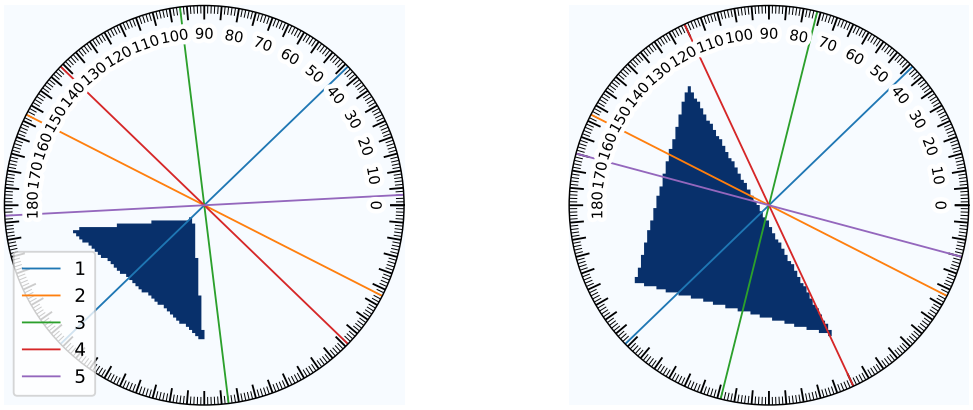


Figure 2.9: The personalized strategies for triangle phantoms achieved by the Actor-Critic policy are demonstrated in these sample results obtained under the end-to-end reward setting.

Table 2.2: Performance comparison of policies on unseen rotations test for TRIANGLES regarding the PSNR values

Policies	Triangles
Learned adaptive policy (end-to-end)	**24.07 ± 2.07**
Learned adaptive policy (increment)	23.78 ± 1.80
Equidistant policy	20.64 ± 1.05

Subsequently, for the right phantom, the agent selects 76°, 115°, and 165° as the following three angles, while for the left phantom, the agent chooses 97°, 136°, and 3°. Notably, these angles are almost tangential to the edges of the triangle phantoms.

We observe that sometimes, the agent selects more angles around the informative angles or repeats its selection when the first two angles are close to the informative angles. This sub-optimal behavior is counter-intuitive and might indicate that the policy network's capacity is too limited to avoid this repetition.

In the context of evaluating unseen rotations across 300 test phantoms, Table (2.2) demonstrates that the Actor-Critic policies outperform the equidistant benchmark. Furthermore, it is observed that the end-to-end reward setting achieves the highest quality in terms of reconstruction.

2.4.6 Experiment 4 - Mixed phantoms with explicit informative angles

In this study, we aim to investigate the capacity of an Actor-Critic agent to recognize and distinguish between different phantoms with varying informative angles and their rotation. Our research methodology trains the agent using a comprehensive dataset (dataset d4) comprising 9,000 phantoms over 300,000 training episodes. In this study, a fixed number of seven angles is employed.

Similar to Experiment 3, shown in Figure (2.10), the training for the mixed phantoms reveals that the incremental reward setting facilitates faster convergence, while end-to-end reward setting results in better performance. Figure (2.11) illustrates the performance of the end-to-end reward setting. It fixes the first two angles to 137° and 46° for the hexagon and triangle, respectively, while it selects 137° and 65° for the pentagon as the first two angles because of some uncertainty for the second angle as mentioned in Experiment 3. The agent then selects the subsequent informative angles based on this prior information by the fixed angles.

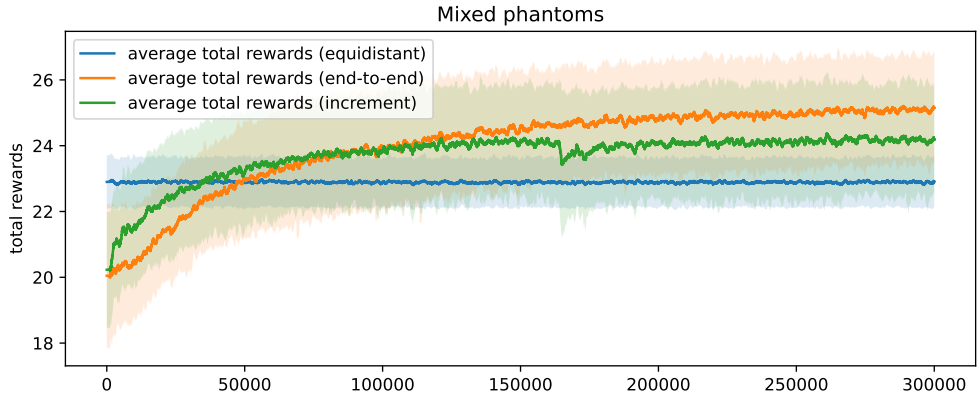


Figure 2.10: This figure compares the performance of Actor-Critic policies, trained on mixed phantoms dataset, with that of an equidistant policy. It displays the training outcomes, with curves showing the mean values and shaded color bands representing the variances.

Drawing from the unseen rotations test conducted on 900 phantoms, as detailed in Table (2.3), it can be observed that the Actor-Critic policies, with both end-to-end and incremental rewards, outperform the equidistant policy.

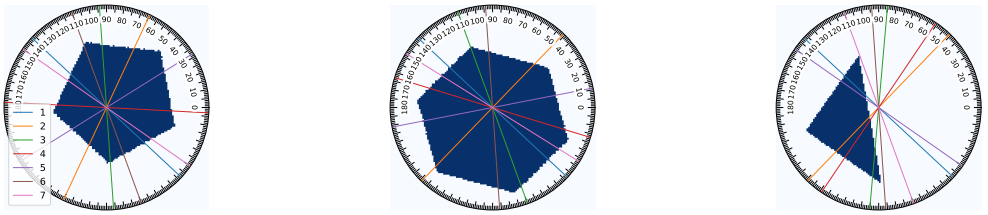


Figure 2.11: The personalized strategies for mixed phantoms achieved by the Actor-Critic policy are demonstrated in these sample results obtained under the end-to-end reward setting.

Table 2.3: Performance comparison of policies on unseen rotations test for MIXED phantoms regarding the PSNR values

Policies	Mixed phantoms
Learned adaptive policy (end-to-end)	**24.85 ± 1.64**
Learned adaptive policy (increment)	24.15 ± 1.55
Equidistant policy	22.94 ± 0.76

2.4.7 Experiment 5 - Generalizability Assessment

To evaluate the generalizability of the proposed method, we apply the trained Actor-Critic policy, initially tailored for mixed phantoms, to various noise levels, phantoms possessing distinct shapes and intensity values. Following this, we train the Actor-Critic policy using simple datasets. Next, to assess its effectiveness, we evaluated how well it could generalize from a simple training dataset to a similar yet more complex dataset.

We investigate the impact of Gaussian noise (5%, 7.5%, and 10%) on the Actor-Critic agent's ability to select informative angles for phantoms. As illustrated in Figure (2.12), the incorporation of Gaussian noise into measurements results in a decrease in performance for both the equidistant and Actor-Critic policies, particularly when compared to outcomes from noise-free measurements. Nevertheless, Actor-Critic policies consistently outperform the equidistant approach, particularly in an end-to-end framework. This indicates that the predominant factors contributing to performance disparities are the difficulties inherent in reconstructing data from noise-contaminated inputs. Comparing the result samples with 5% Gaussian noise from end-to-end rewards in Figure (2.13) to those in Figure (2.11) from Experiment 4, we find that the noise in measurements has a substantial influence on the angle selection strategy, including the fixed angles and the informative angles selection orders afterward. To better understand the differences between the two policies in Figures (2.11) and

Table 2.4: Performance comparison of policies on test for phantoms with different shapes regarding the PSNR values

Policies	Rectangular	Heptagon	Octagon
Learned adaptive policy (end-to-end)	21.28 ± 2.66	23.08 ± 0.74	23.19 ± 1.24
Learned adaptive policy (increment)	21.89 ± 2.52	**23.32 ± 0.57**	23.46 ± 0.90
Equidistant policy	**23.66 ± 1.20**	22.75 ± 2.16	**24.24 ± 0.43**

(2.13), we show the policy results for triangles in Figure (2.14). Our analysis reveals that the policy for clean data is tightly clustered around informative angles, whereas the policy for noisy data is more broadly distributed. We also observe that the policy realizes adaptive angle selection, where the probability of a chosen angle decreases significantly, followed by an increase in the probability of some angles with a small probability.

In the test involving unseen rotations, conducted on 900 phantoms with variable noise levels as depicted in Figure (2.15), the Actor-Critic policies consistently surpass the equidistant policy, with the end-to-end reward setting maintaining the most superior performance across the three distinct noise levels. This suggests again that the observed performance variation predominantly stems from the reconstruction challenges posed by noisy measurements rather than the efficiency of the trained policy itself. Additionally, models trained on clean measurements generally exhibit enhanced resilience to noisy measurements compared to those trained on noisy datasets, suggesting that training with noisy data might hinder the optimal performance of the Actor-Critic policies.

In terms of shape variations, we analyze phantoms with fewer edges, such as rectangles, and those with a greater number of edges, specifically heptagons, and octagons. For each shape category, we have studied a sample size of 900 phantoms. As deduced from Table (2.4), while the Actor-Critic policy demonstrates a degree of generalizability for diverse phantom shapes, such as heptagon, its efficacy is somewhat restricted. The policy with an incremental setting demonstrates enhanced generalizability.

For variations in intensity values, we assess phantoms (dataset d4) with four distinct intensity ranges: [0.9,1), [0.7,1), [0.5,1), and [0.1,1). For each of these intensity categories, our study encompassed a sample size of 900 phantoms. Insights drawn from Table (2.7) suggest that as the intensity value range expands, the generalizability of the Actor-Critic policy diminishes. However, the policy employing an end-to-end setting showcases superior generalizability.

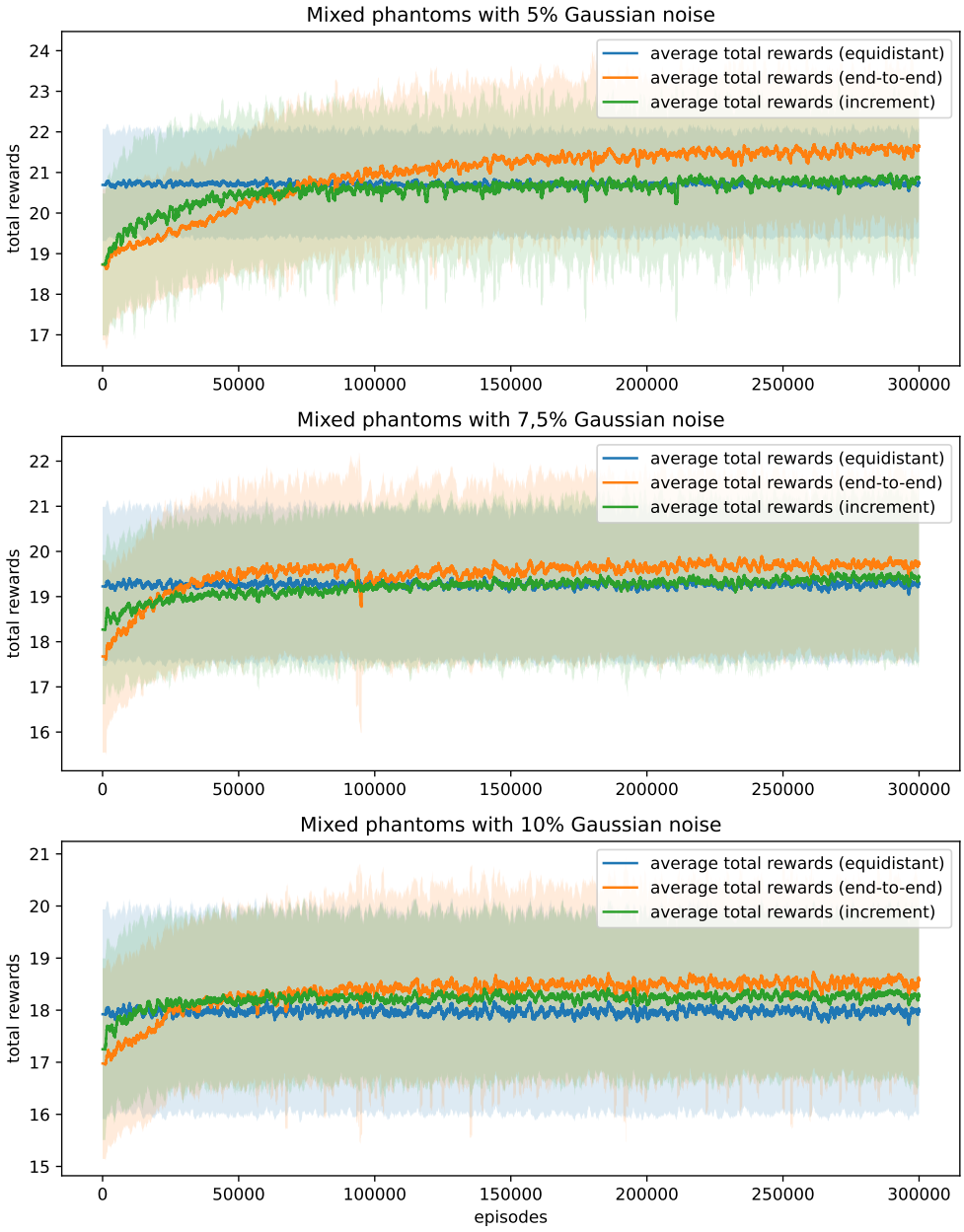


Figure 2.12: This figure compares the performance of Actor-Critic policies, trained on a mixed phantoms dataset, with that of an equidistant policy. It displays the training outcomes while considering the impact of Gaussian noise (5%, 7.5%, and 10%), with curves indicating the mean values and shaded color bands representing the variances.

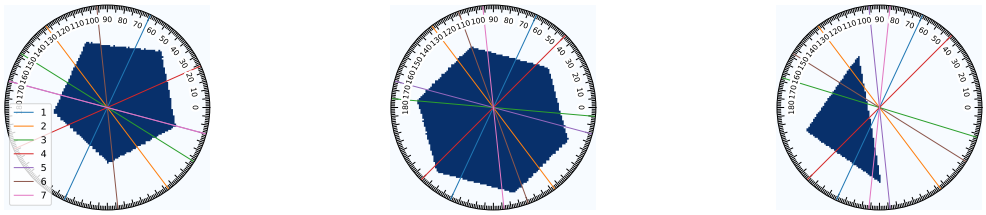


Figure 2.13: The personalized strategies for mixed phantoms with 5% Gaussian noise on measurements achieved by the Actor-Critic policy are demonstrated in these sample results, obtained under the end-to-end reward setting.

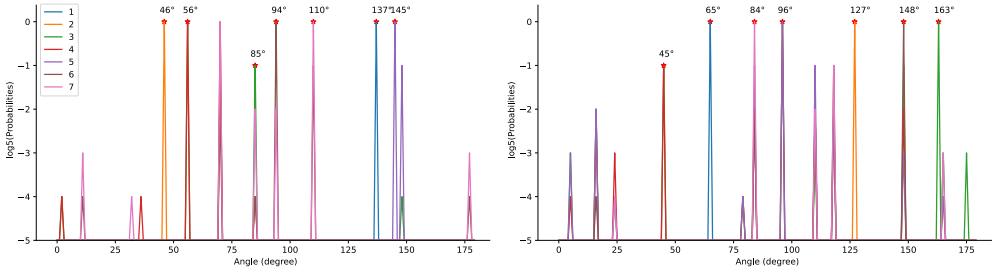


Figure 2.14: Comparison of policies for triangles. The policy on the left is trained on clean data, while the policy on the right is trained on noisy data. The probabilities are scaled by $\log 5$ to magnify the small probabilities. The red star on each colored line represents the selected angle based on the corresponding probability distribution. Once an angle is selected, its probability becomes zero or is decreased for subsequent selections.

Finally, we initially trained the Actor-Critic policy on a comprehensive dataset that includes all phantoms from categories d1), d2), and d4). This approach aimed to develop a versatile and adaptable policy capable of handling a variety of shapes and complexities. With a set configuration of seven angles across 300,000 training episodes, Figure 2.16 showcases the training results for the Actor-Critic policies and uses the equidistant policy as a comparative baseline. The Actor-Critic policies surpass the equidistant policy in performance, though the margin is modest. This narrow performance gap can be attributed to the specific nature of the training datasets, particularly the circle and ellipse phantoms. Circles possess uniformly informative angles, while ellipses offer nearly uniform informative angles, as shown in Experiment 2, considering the seven angles. Interestingly, the performances of Actor-Critic policies, whether configured for end-to-end or incremental learning, show a remarkable similarity.

Subsequently, to assess the ability of the policy to generalize from a simpler training

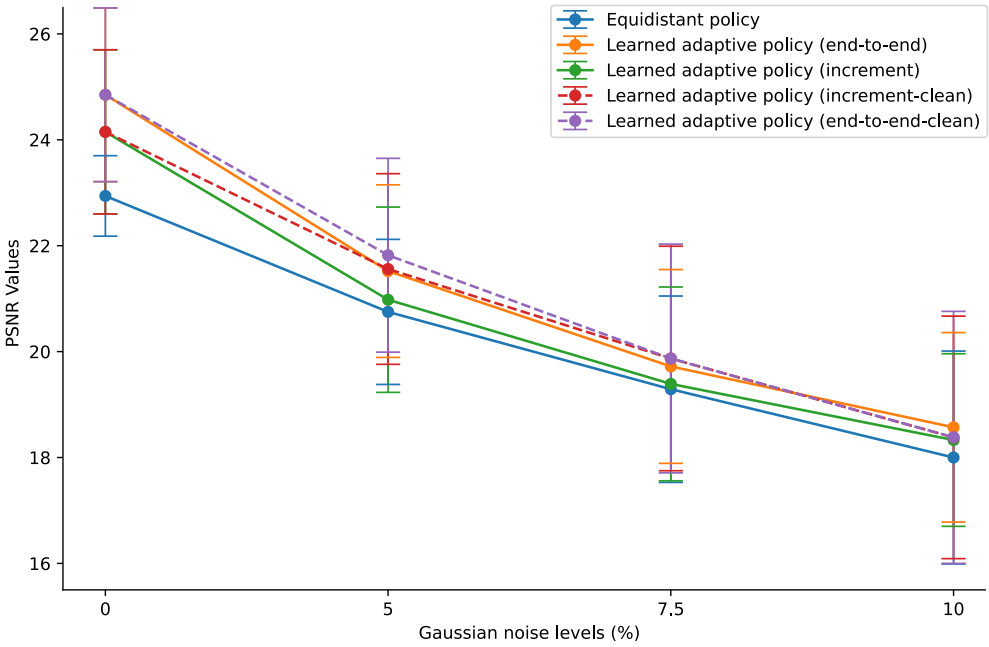


Figure 2.15: Comparison of policies considering different Gaussian noise levels for the mixed dataset d4): the results demonstrate the unseen rotations outcomes over. The graph depicts the efficacy of learned adaptive policies, both in incremental and end-to-end settings, against the equidistant policy. Notably, both clean and noisy training conditions are considered for the adaptive policies. The x-axis represents the Gaussian noise percentage, while the y-axis captures the performance metric. Error bars indicate the variability in performance at each noise level.

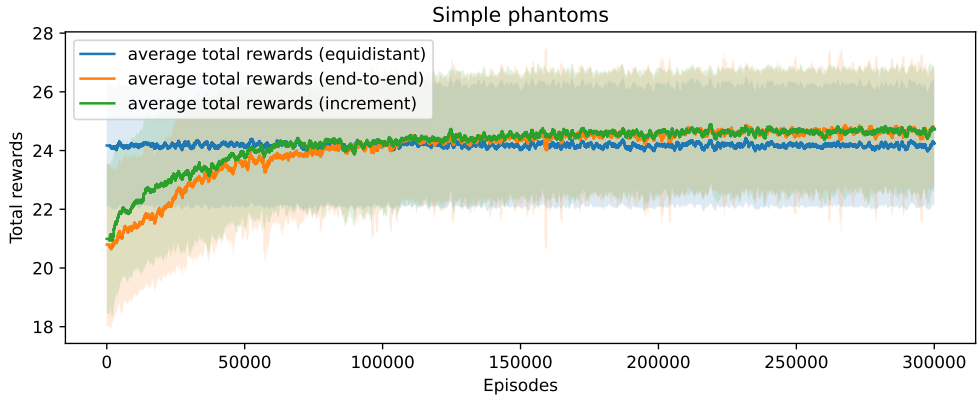


Figure 2.16: This figure compares the performance of Actor-Critic policies, trained on simple phantoms from datasets d1), d2), and d4), with that of an equidistant policy. It displays the training outcomes, with curves showing the mean values and shaded color bands representing the variances.

Table 2.5: Policy Performance on Complex Combined Phantoms: A comparison of policies trained on simple phantoms, evaluated on complex combined phantoms (d5) using PSNR values.

Policies	**Combined phantoms d5)**
Learned adaptive policy (end-to-end)	23.27 ± 1.43
Learned adaptive policy (increment)	23.20 ± 1.67
Equidistant policy	23.24 ± 1.25

set to a more complex one, we evaluate the trained Actor-Critic policy on datasets d5) and d6) with a size of 900 phantoms, respectively. This category encompasses a mix of two distinct shape types found in the training dataset, thereby providing a challenging test for its adaptability. The results for the dataset d5), depicted in Table (2.5), indicate that the performance of these three policies is quite similar due to the inclusion of round shapes. In contrast, the results for dataset d6), shown in Table (2.6), demonstrate the superiority of both Actor-Critic policies. These results highlight the Actor-Critic's proficiency in handling increased complexity.

Table 2.6: Policy Performance on Complex Combined Phantoms: A comparison of policies trained on simple phantoms, evaluated on complex combined phantoms (d6) using PSNR values.

Policies	Combined phantoms d6)
Learned adaptive policy (end-to-end)	22.35 ± 1.12
Learned adaptive policy (increment)	**22.49 ± 1.27**
Equidistant policy	22.20 ± 0.54

Table 2.7: Performance comparison of policies on test for phantoms with different intensity values regarding the PSNR values

Policies	[0.9,1)	[0.7,1)	[0.5,1)	[0.1,1)
Learned adaptive policy (end-to-end)	**23.91 ± 1.86**	**23.30 ± 2.35**	**22.51 ± 2.91**	21.21 ± 3.04
Learned adaptive policy (increment)	22.41 ± 2.98	21.41 ± 3.73	18.56 ± 3.92	17.29 ± 3.42
Equidistant policy	21.77 ± 0.77	21.43 ± 0.76	21.36 ± 0.74	**21.31 ± 0.71**

2.4.8 Experiment 6 - Complex phantoms with implicit informative angles

This study seeks to evaluate the effectiveness of the Actor-Critic agent when applied to complex, multi-material phantoms that more closely resemble realistic scenarios. The agent is trained using dataset d7) to examine its capability in a-posteriori adaptation and on dataset d8) to evaluate the impact of static components on a-priori angle selection.

The number of phantoms used for training is 3,000 and the episodes for training are 150,000. The training outcomes during the final 2,000 episodes are illustrated in Figure (2.17) for dataset d7) and Figure (2.18) for dataset d8). Both figures affirm the superior efficacy of the Actor-Critic policies. As the number of angles increases, the differences in results among the three policies become less pronounced. This convergence suggests that with a sufficient number of angles, even an equidistant policy can achieve high-quality reconstructions. Noteworthily, the Actor-Critic policy under the end-to-end reward configuration still manifests optimal performance.

In Figures (2.19) and (2.20), we showcase the training results for two different configurations: five angles in group a) and nineteen angles in group b). The distinction in angle choices for these configurations is emphasized by the red lines in both groups a) and b), illustrating the Actor-Critic's proficiency in a-posteriori adaptation. These red lines delineate the Actor-Critic agent's ability to monitor the rotations of the ellipses. To understand the impact of the static component on a-priori angle-selection,

we juxtapose the outcomes from datasets d7) and d8) as seen in Figure (2.21). In our analysis of the final 2000 episodes, angles selected over 1750 times are classified as belonging to the most frequently selected category. In comparison to group a), group b) incorporates an increased count of angles, characterized by a denser distribution, attributable to the introduction of an additional static ellipse. Collectively, these insights emphasize the Actor-Critic agent's proficiency in managing both a-priori angle selection and a-posteriori adjustments.

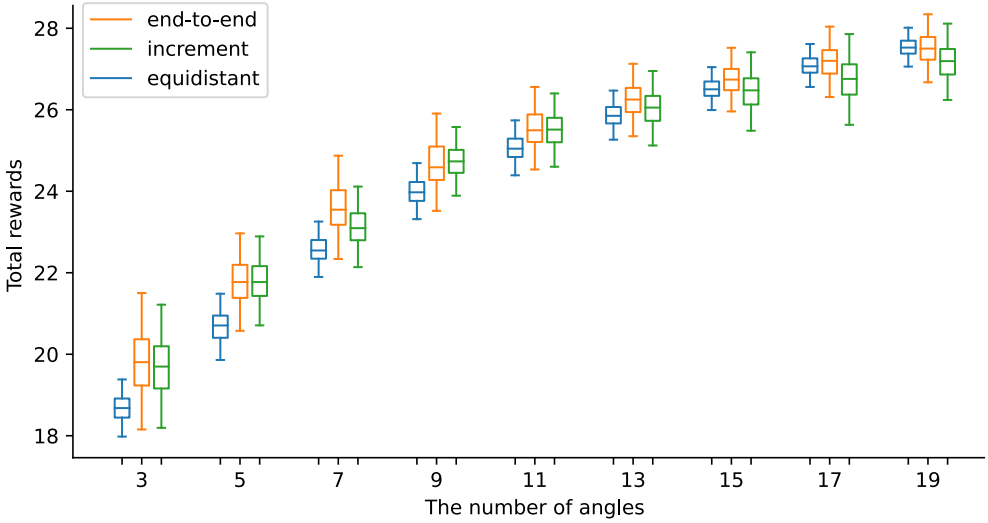


Figure 2.17: Comparison of policies considering different numbers of angles for the complex dataset d7): the results demonstrate the training outcomes over the last 2,000 episodes. The box represents the interquartile range in these plots, spanning from the first to the third quartile of the data distribution. The median value is displayed as a line within the box. The whiskers extend from the box to illustrate the range of the data distribution beyond the interquartile range.

2.5 Discussion

The results indicate that for both categories of phantoms—those with clear informative angles and those without—the reinforcement learning policy are able to achieve superior performance compared to the uninformed, equidistant policy. Empirical evidence suggests that the reinforcement learning policy can devise a strategy for the static component based on a-priori knowledge while authentically implementing a-posteriori adaptations. This complements the findings from [78], whose numerical studies could

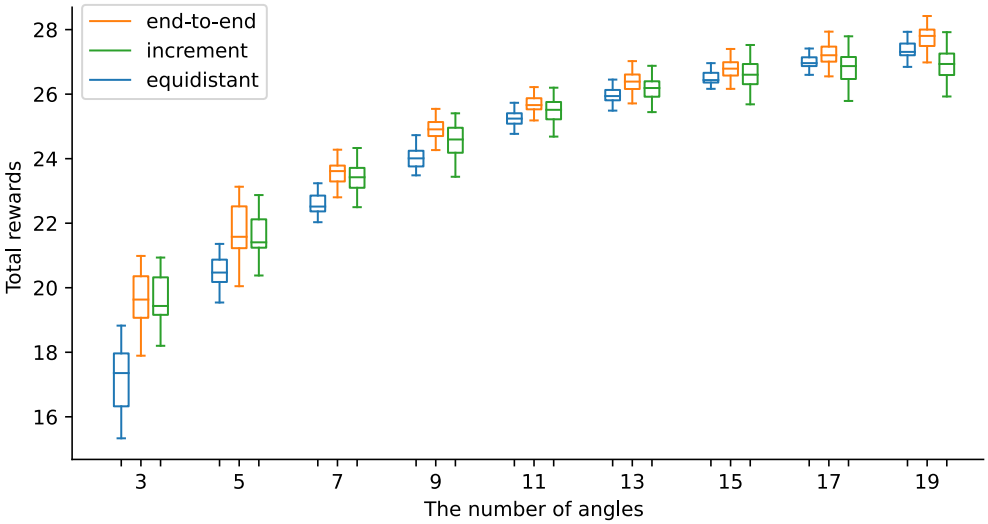


Figure 2.18: Comparison of policies considering different numbers of angles for the complex dataset d8): the results demonstrate the training outcomes over the last 2,000 episodes. The box represents the interquartile range in these plots, spanning from the first to the third quartile of the data distribution. The median value is displayed as a line within the box. The whiskers extend from the box to illustrate the range of the data distribution beyond the interquartile range.

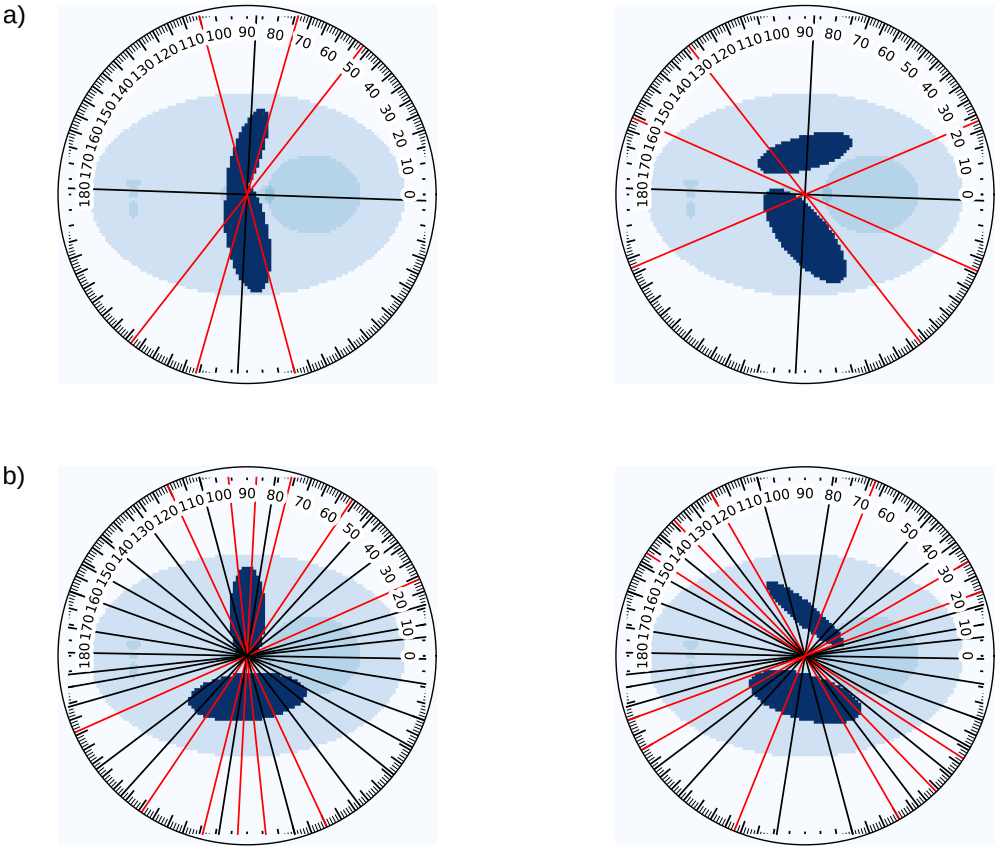


Figure 2.19: Illustration of personalized strategies for complex phantoms in d7) achieved using the Actor-Critic policy under the end-to-end reward setting. Two groups are depicted based on the number of angles: a) five angles; b) nineteen angles. Rows one and two present a comparison of distinct phantoms, with red lines emphasizing the a-posteriori adaptation due to ellipse rotations.

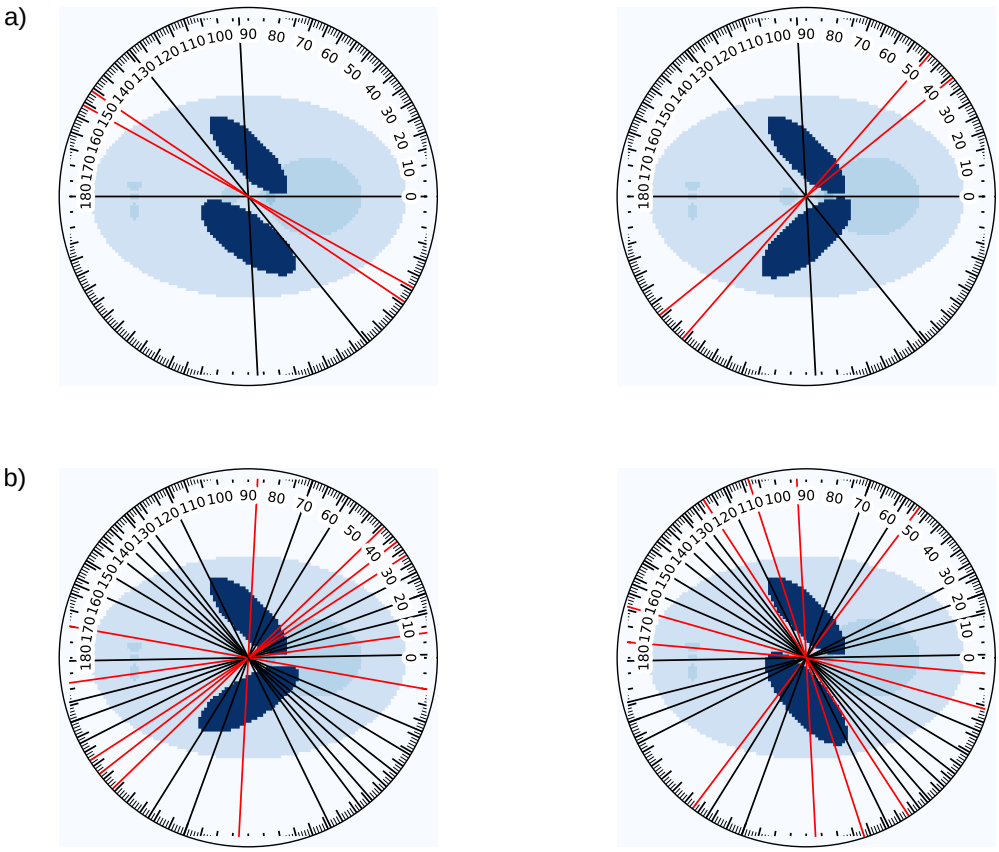


Figure 2.20: Illustration of personalized strategies for complex phantoms in d8) achieved using the Actor-Critic policy under the end-to-end reward setting. Two groups are depicted based on the number of angles: a) five angles; b) nineteen angles. Rows one and two present a comparison of distinct phantoms, with red lines emphasizing the a-posteriori adaptation due to ellipse rotations.

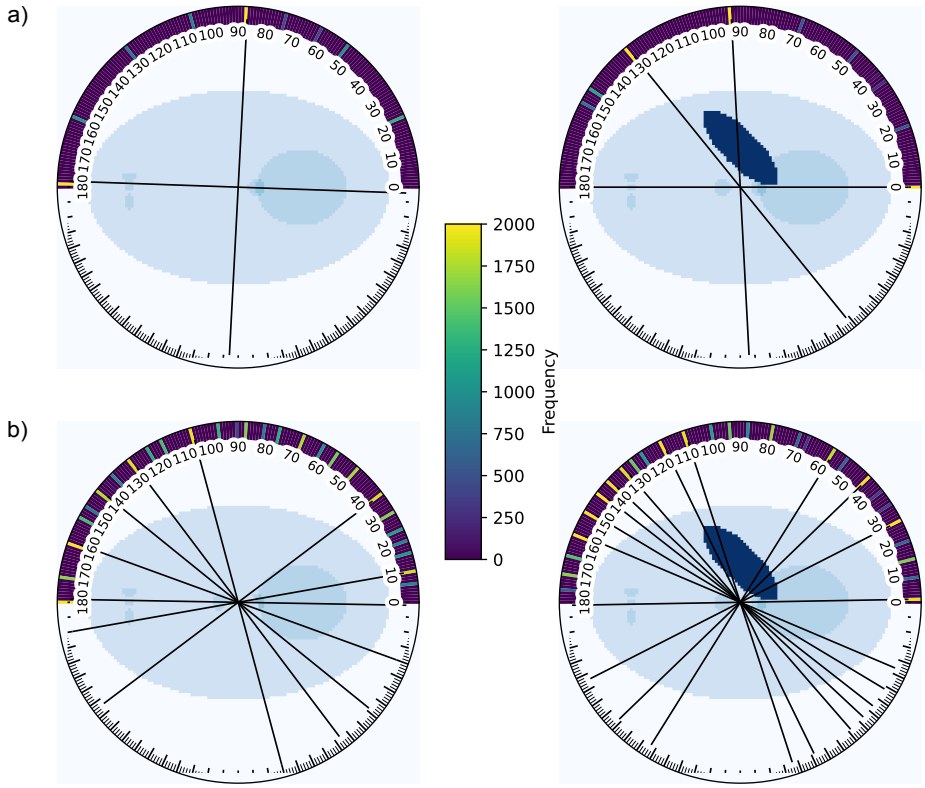


Figure 2.21: The figure illustrates the distribution of angles most commonly chosen, emphasizing intrinsic preferences devoid of dynamic interferences. The depicted phantoms signify the omission of any dynamic element. The gradation in color is indicative of the frequency with which specific angles were chosen over the last 2000 episodes, with each angle correlating to a color on the provided scale. Notably, angles that fall within the top frequency bracket (1750 to 2000 occurrences) are accentuated by pronounced black lines.

not answer this important question. Furthermore, our approach simplifies the extension to other design parameters beyond angle selection. While angle selection involves straightforward sampling from one dimension of the complete dataset, altering other design parameters, such as zooming into a region of interest, introduces more complex changes in data geometry, like resolution adjustments. Designing a network to operate effectively within such a variable data environment presents significant challenges. However, in our approach, the reconstruction method intuitively accommodates these variations and seamlessly converts them to a consistent image space format, thus alleviating the complexities inherent in handling variable data geometries.

Importantly, the trained reinforcement learning policies exhibit generalization capabilities on the test dataset, including rotations not encountered during training. Introducing varying levels of noise in measurements diminishes the potential performance improvements across all policies. Yet, reinforcement learning policies continue to exhibit superior performance. This suggests that the primary challenge arises from the reconstruction of noisy data. Although the method's generalizability is somewhat limited when faced with diverse shapes and intensity values, it remains relevant under particular scenarios, especially when intensity values mirror those encountered during training. Additionally, initially trained using simpler phantoms, these policies are later evaluated on more complex ones, essentially aggregations of the simpler phantoms. This illustrates the policies' capability to adapt from basic to more intricate geometries. We have also determined that a robust training dataset is crucial. Fortunately, high-quality datasets like CAD models are accessible for industrial applications. The insights from Experiment 5 highlight the flexibility of the reinforcement learning policy in adapting to more complex or varied situations, suggesting room for further improvement.

In addition, we conducted numerical experiments to compare end-to-end and incremental reward functions. The end-to-end reward function achieves the highest average performance on both the training and test datasets. This indicates its effectiveness in guiding the reinforcement learning agent toward optimal solutions. On the other hand, the incremental reward function demonstrates faster convergence during training. In the future, we will investigate further how to design reward functions that share both of these desirable properties.

In the future, our work can be extended in the following ways: Firstly, instead of using SIRT as an image reconstruction method, we will use deep learning-based reconstruction methods, trained end-to-end. The integration of DL in the reconstruction process is anticipated to significantly expedite and enhance the efficiency of our

image reconstruction pipeline. This adaptation is expected to render the process more efficient, particularly aligning it with the demands of real-time application scenarios. Conceptually, the current separation between image reconstruction and RL involves different a-priori assumptions about the class of images being analyzed. By developing a coherent framework that unifies these assumptions for both tasks, we expect to achieve more consistent and reliable outcomes. Secondly, our forthcoming research will involve rigorous testing of our method against datasets containing defects to validate its effectiveness. Working towards this integration by incorporating defects into our phantom studies and factoring defect detection into the cost function. Thirdly, we will focus on improving the policy network to mitigate the issue of repeated angle selections. Fourthly, we restricted ourselves to a simple 2D parallel-beam geometry to obtain scenarios in which optimal angle selection strategies are known, and the results of trained policies can be interpreted more easily. In the future, we will extend the approach to more complex and realistic 3D geometries with additional degrees of freedom, such as tilting and zooming. Fifthly, we will continue exploring and strengthening the generalization capabilities of our approach. This enhancement is crucial for ensuring that the algorithm can adapt effectively to a wide range of imaging situations. Finally, a significant aspect of our future work will involve testing and validating our algorithm on actual CT scans. This step is fundamental to transition from theoretical and simulated environments to real-world applications, and will be a critical area of focus in our continued research.

2.6 Conclusion

Compared to classical, computationally prohibitive approaches to solve the sequential OED problem of adaptive angle selection in X-ray CT, deep reinforcement learning avoids direct gradient computation on the high-dimensional, non-convex, bi-level optimization problem. Instead, it learns non-greedy strategies to solve it for a particular class of phantoms during an offline training phase which can then be applied fast and efficiently online to scans of new phantoms. We posed the sequential OED problem as a POMDP and utilized the Actor-Critic network combining a shared encoder network to learn an optimal policy. In our numerical studies with 2D CT scenarios mimicking industrial, in-line CT inspection, we could demonstrate that our approach learns efficient, truly adaptive policies that achieve better performance in terms of reconstruction quality. We introduced two different reward function settings, namely, the end-to-end and incremental reward settings. Both settings lead to stable learning

processes, consolidating reinforcement learning as a reliable and extremely promising method for sequential OED. To conclude, our work demonstrates the potential of using reinforcement learning for solving sequential OED problems in inverse problems and imaging - in particular to automate angle selection and improve CT imaging efficiency, providing a flexible and adaptive approach for various CT imaging scenarios in the Industry 4.0.

Chapter 3

Task-Adaptive Angle Selection for Computed Tomography-Based Defect Detection

3.1 Introduction

X-ray Computed Tomography (CT) plays a pivotal role in non-destructive quality control, yet it remains predominantly an off-line tool due to the trade-off between scan speed and scan quality. Standard CT protocols use evenly spaced angles around the object, typically requiring a large number of angles to ensure quality. However, in particular in sparse angle CT, the geometry of the object under analysis heavily influences which angular projections carry more or less useful information for the specific task to be performed, e.g., for defect detection [46, 29]. Optimal Experimental Design (OED) aims to identify the optimal scanning angles to achieve a given task. We discern

This chapter is based on:

Tianyuan Wang, Virginia Florian, Richard Schielein, Christian Kretzer, Stefan Kasperl, Felix Lucka, and Tristan van Leeuwen. "Task-Adaptive Angle Selection for Computed Tomography-Based Defect Detection". In: *Journal of Imaging* 10.9 (2024), p. 208. DOI: 10.3390/jimaging10090208

two primary categories: static OED and sequential OED. Static OED determines all experimental parameters in advance, without incorporating feedback from subsequent observations. In contrast, sequential OED relies on continuous feedback from each stage of the experiment, allowing for adjustments and adaptations based on emerging data. In the context of enhancing the speed of in-line CT inspections, the development of sequential OED for adaptive angle selection strategies is essential. Such a strategy would allow for a more focused scanning process, dynamically adjusting the angles and reducing the number of necessary angles, tailored to meet the specific needs of the inspection task. This approach promises to streamline the scanning process, making it quicker and more responsive to the unique characteristics of each inspection scenario.

Several researchers have employed prior information in the form of generic statistical models to determine the most informative angles using approaches based on OED [9, 22, 14, 39]. However, computational challenges have limited the application of such methods to off-line settings. In industrial applications, prior information in the form of Computer-aided design (CAD) models are often available. Therefore, many researchers utilized them for off-line training to determine the informative angles. The resulting predefined trajectories can then be rapidly applied in real settings. Fischer *et al.* [29] proposed a CAD-based method to optimize task-specific trajectories using a detectability index [82]. The index is computed by evaluating trajectory fitness against a user-defined frequency template, utilizing a modulation transfer function and the noise power spectrum. There are some extended works based on this: Herl *et al.* [42] considered a Tuy-based metric to guarantee the completeness of the data acquisition. Schneider *et al.* [69] used the regressive ResNet-34 to predict this index for each angle. Then, they used an integer optimization method to generate the trajectory based on these predictions. Matz *et al.* [54] proposed geometry-based metrics using a discrete wavelet transform to find which angle can capture more information about edges. Bussy *et al.* [94] employed the Q-Discrete Empirical Interpolation Method to obtain a subset from full angles. Furthermore, they introduced additional constraints considering the ray's attenuation [15].

The need for adaptability in in-line use, particularly regarding angle selection, is underscored by the fact that these offline-optimized methods, while efficient in practical applications, often require precise alignment with actual scans [10]. To address this, we explored the potential of a Deep Reinforcement Learning (DRL) approach in the development of the sequential OED for adaptive angle selection in [97]. So far, we only considered approaches that choose a fixed number of angles to optimize generic image reconstruction quality measures.

This paper extends our previous work [97] and addresses challenges of identifying defects in products using in-line CT. The contributions of this study include:

1. Incorporating prior knowledge about potential defects into the DRL reward function to guide task-specific angle selection for defect detection;

2. Acknowledging that defects in industrial products are infrequent, and that defective and non-defective samples may require different numbers of angles, our method introduces adaptability in angle count. This flexibility allows for a more detailed inspection when defects are suspected, enhancing defect detection while simultaneously optimizing the use of scanning resources;

3. Incorporating task-specific, such as image contrast or segmentation quality in the training reward. This integration of DRL with defect detection strategies significantly enhances both the rapidity and precision of defect identification, demonstrating the feasibility of an automated and efficient process for task-specific angle selection in CT scans.

The paper is structured as follows: The following section introduces the basic concepts and notation we use for the inverse problem in CT imaging and Reinforcement Learning (RL) for sequential OED. Section 3.3 elaborates on DRL and defect detection methodologies and outlines the integration of defect information into the RL reward function. The fourth section presents three experiments validating the algorithm. The fifth section discusses the advantages, limitations, and prospects for future research. The paper concludes in the final section with a summary of our findings.

3.2 Background

3.2.1 Inverse problem of CT

In sparse-angle CT, we aim to reconstruct $\bar{\boldsymbol{x}}$ from a small number of angles $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$, which presents a underdetermined and ill-posed inverse problem, which we will denote as

$$\boldsymbol{y}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\bar{\boldsymbol{x}} + \boldsymbol{\epsilon}(\boldsymbol{\theta}), \tag{3.1}$$

Traditional approaches like Filtered Back-Projection (FBP) are inadequate here, as they rely on a complete range of angles. We will use the Simultaneous Iterative Reconstruction Technique (SIRT) [27], which obtains a reconstruction by solving the

constrained weighted least-squares problem

$$\widehat{\boldsymbol{x}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{x}\in\Lambda} \|\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x} - \boldsymbol{y}(\boldsymbol{\theta})\|_R^2 , \tag{3.2}$$

using a pre-conditioned projected gradient-descent scheme:

$$\boldsymbol{x}^{(k+1)} = \Pi_\Lambda \left(\boldsymbol{x}^{(k)} + C\boldsymbol{A}(\boldsymbol{\theta})^T R \left(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}^{(k)} - \boldsymbol{y}(\boldsymbol{\theta}) \right) \right) . \tag{3.3}$$

In the above, Π_Λ denotes a projection onto the constraint set (e.g., bound constraints $a \leq x_i \leq b$), the diagonal matrix R contains the reciprocals of row sums of $\boldsymbol{A}(\boldsymbol{\theta})$ and the diagonal matrix C contains the column sums of $\boldsymbol{A}(\boldsymbol{\theta})$.

3.2.2 Reinforcement learning for sequential OED

OED is a methodological framework developed to select experimental designs that maximize the information gained about parameters of interest [64]. In the example of (3.1), OED would aim to choose a set $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$ such that $\boldsymbol{y}(\boldsymbol{\theta})$ contains maximal information about a feature of $\bar{\boldsymbol{x}}$. In sequential OED, as an extension, the experimental designs can be adapted sequentially based on the outcomes of previous experiments, e.g., we can choose each new angle θ_k based on the measurements $\boldsymbol{y}(\{\theta_1, \ldots, \theta_{k-1}\})$.

While conventional OED formulations lead to complex optimization problems over utility functions measuring the information gain, we are considering using Reinforcement Learning (RL) to solve the sequential OED problem. RL is a machine learning paradigm in which intelligent agents learn to take actions in a dynamical environment that will maximize a reward they obtain [84]. RL is typically grounded in Markov Decision Processes (MDP), a framework comprising action and state spaces, transition, and reward functions. RL algorithms aim to learn a policy —a strategy for selecting actions based on the current state —that maximizes cumulative rewards over time. Partially observable MDP (POMDP) extends this framework to scenarios, where the current state may not always be fully observable, and the agent has to construct a belief state that its actions are then based on.

We utilize RL within the POMDP framework to address the sequential OED problem following [77]. The RL algorithm operates by mapping the belief state at step t, denoted as $\boldsymbol{s}_t$, to a probability distribution over the possible experimental designs or actions. The action for the next experiment, a_t, is then chosen based on this probability distribution. The objective is to learn a policy $\pi(a_t|\boldsymbol{s}_t)$ that optimizes the utility

function of the sequential OED, essentially guiding the experimental design process in a way that maximizes the overall gain from the experiments conducted. Consequently, this methodology allows for the optimization of policy parameters rather than the design parameters themselves.

In [97], we described a POMDP model of the sequential OED problem of adaptively selecting a fixed number of scan angles in X-ray CT in terms of states, actions, and rewards associated with state transitions. The DRL training in this setup was focused on developing an optimal policy aimed at maximizing total rewards, primarily enhancing the reconstruction quality by choosing the most suitable actions based on the current state. The components of this model include:

- *Observation space*: it consists of a set of measurements $\boldsymbol{y}$ generated according to Equation 3.1.
- *State space*: we consider the reconstruction $\widehat{\boldsymbol{x}}$ of the underlying ground truth $\bar{\boldsymbol{x}}$ as a belief state, which can be obtained via SIRT (3.3).
- *Action space*: it consists of 360 integer angles $\boldsymbol{\theta} = \{\theta_0, \theta_1, ..., \theta_{359}\}$ from the range [0, 360).
- *Transition function and observation function*: the transition function is deterministic, as the underlying ground truth remains unchanged. On the other hand, the data model given by the forward model serves as the observation function, from which we only consider measurement samples.
- *Reward function*: The reward function primarily evaluated the reconstruction quality using the Peak Signal-to-Noise Ratio (PSNR) value.

3.3 Methods

3.3.1 Task-adaptive angle selection for defect detection

Building on the approach outlined above, we now introduce a more complex reward function to facilitate adaptive, task-specific angle selection with a flexible number of angles. We propose an integrated pipeline where angle selection aids in improving defect detectability. As shown in Figure 3.1, the pipeline involves training two key components: the DRL network and the evaluation for defect detectability. The DRL network follows the architecture described in our previous work [97]. It includes a deep neural encoder network that extracts informative features from reconstructions at

each step. Additionally, it incorporates an Actor-Critic network to estimate the value function and output the probability distribution over the action space. The bottleneck layer of the encoder network serves as the input to the Actor-Critic network.

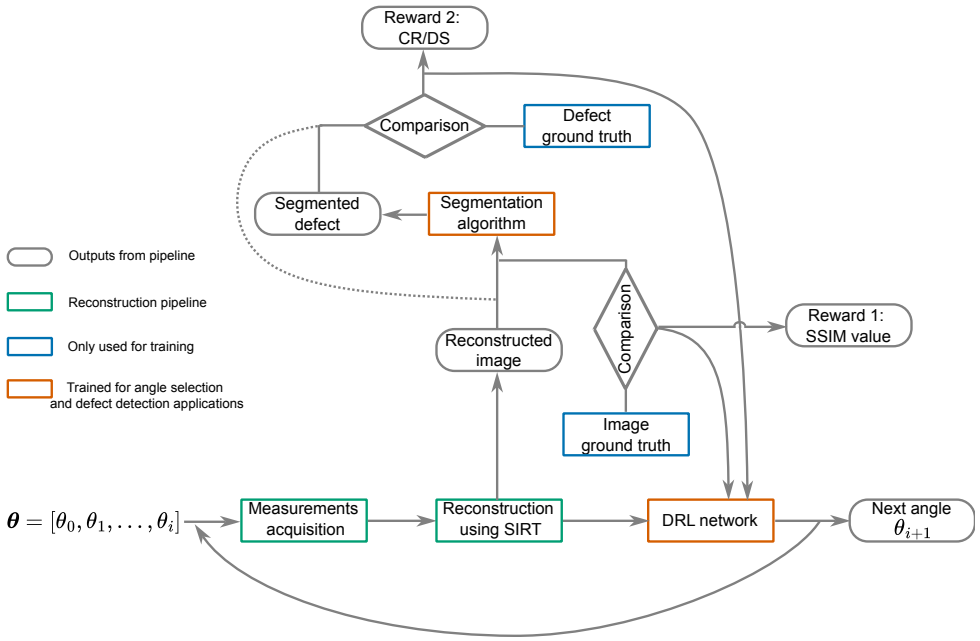


Figure 3.1: The DRL network selects the next angle θ_{i+1}, which is used, along with the previously selected ones and measurements, to gather a new reconstruction. The SSIM is then calculated to assess the similarity between this reconstruction and the ground truth image. Concurrently, the contrast ratio value is derived from the ground truth of the defect (indicated by the dashed line); alternatively, the dice score value is determined by comparing the defect segmentation output with the defect's ground truth (depicted by the solid line). These metrics inform the DRL network's decision-making for further angle selection.

The pipeline operates as follows: To explain one step of this sequential pipeline, we assume a set of angles $\boldsymbol{\theta} = \{\theta_0, \theta_1, ..., \theta_i\}$ and their corresponding measurements are acquired. The SIRT algorithm with constraints is then used for image reconstruction. This reconstructed image serves as the input for the DRL network.

The reward function considers reconstruction quality, which is measured by comparing the reconstruction to the ground truth, often using the Structural Similarity (SSIM) index. It also evaluates defect detectability, described by metrics such as the Contrast Ratio (CR) or Dice Score (DS) for segmentation precision. A negative reward is assigned at each step to encourage the use of fewer angles. More details about

this reward function are provided in the following section.

The policy maps the input reconstructed image to a probability distribution over the action space, from which the next angle is selected by sampling. The policy is trained to maximize the total rewards. The computational costs spend on the training process. While the training processis computationally intensive, the trained model can be applied quickly in real-world applications.

3.3.2 Reward function

The reward function considers two primary objectives: defect detectability and reconstruction quality. The detectability of defects, given the current reconstruction $\widehat{\boldsymbol{x}}_{k+1}$ and the ground truth of the defect $\bar{\boldsymbol{z}}$, will be assessed by a measure $D(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{z}}) \in [0,1]$, where higher values mean a better detectability. We will describe two examples of $D(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{z}})$ in more detail in the following section. To asses reconstruction quality, we will use the SSIM index instead of the PSNR value, because the SSIM also takes values in the range [0,1]. With this, the reward function is defined as:

$$R(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}, \bar{\boldsymbol{z}}) = \begin{cases} \alpha(SSIM(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) + D(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{z}})) & \text{if stop criteria is reached} \\ -a & \text{otherwise} \end{cases} \tag{3.4}$$

The stop criteria are reached if the number of angles $k+1$ is equal to a maximum M, or if the reconstruction quality $SSIM(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}})$ is above a threshold Q. The rationale behind setting M is to prevent the process from selecting an excessively large number of angles, while the threshold Q ensures that angle selection stops in defect-free cases, where $D(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{z}}) = 0$ at each step, and that the reconstruction quality is high enough for reliable defect determination. The term α is a scale. The term $-a$ acts as a penalty at each step, encouraging the RL agent to select fewer angles to meet the stop criteria. Both the Contrast Ratio (CR) metric and Dice Score (DS) metric can serve as the component $D(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{z}})$ of the reward function (3.4).

To effectively assess the detectability of defects, it is important to introduce specific metrics that quantify how clearly a defect is visible. The CR measures the visibility of a defect against the background material in terms of the average intensity difference:

$$CR = \frac{|\mu(I_{\text{defect}}) - \mu(I_{\text{background}})|}{\mu(I_{\text{background}})} \tag{3.5}$$

In this formula, $\mu(I_{\text{defect}})$ represents the mean intensity value of the defect area, while

$\mu(I_{\text{background}})$ denotes the mean intensity value of the surrounding background area. A greater CR value between a defect and the surrounding material often correlates with better detectability.

In practical terms, however, one must also take into account the actual segmentation of the defect. To this end, we explore unsupervised defect segmentation using the K-means clustering algorithm, a method celebrated for its uncomplicated yet effective approach to categorizing complex data into distinct clusters.

K-means clustering begins by determining a predetermined number of clusters, known as K, with each cluster represented by a central value or 'centroid'. The algorithm then proceeds in an iterative fashion, assigning each data point to the closest centroid using a distance metric, typically Euclidean distance. Centroids are recalculated in each iteration, and this process repeats until the centroid positions no longer change significantly.

Our paper leverages the K-means clustering algorithm to identify defects within material samples. By initializing the algorithm to create two clusters, we effectively distinguish between normal and defective regions. The defective cluster is then used to guide the manual segmentation of defects, illustrating the algorithm's practical application in materials analysis.

When this semantic segmentation method is applied to the reconstructed image, one can also consider a metric that directly measures the accuracy of the segmented image: The DS compares the predicted segmentation of defects against the actual (ground truth) segmentation. It is expressed as:

$$DS = \frac{2 \times |P_{\text{predicted}} \cap P_{\text{ground truth}}|}{|P_{\text{predicted}}| + |P_{\text{ground truth}}|} \tag{3.6}$$

Here, $P_{\text{predicted}}$ refers to the set of pixels identified as defects in the predicted segmentation and $P_{\text{ground truth}}$ represents the set of pixels identified as defects in the actual ground truth segmentation. Thus, $P_{\text{predicted}} \cap P_{\text{ground truth}}$ is the set of correctly identified as defect pixels. The cardinality operator $|\cdot|$ measures the size of the sets, i.e., the number of active pixels.

These metrics offer insights into how effectively the chosen imaging angles can accentuate defects in the scans. By incorporating CR or DS into the reward function, we can systematically evaluate and optimize the angle selection process, ensuring that it maximizes the visibility and detectability of defects.

3.4 Results

In the results section of our paper, we explore the efficacy of the pipeline in selecting informative angles that enhance defect detectability. We compare our RL-based method to non-adaptive angle selection criteria. These non-adaptive angle selection criteria use the same number of angles, which is determined by the RL-based method. Our experimental analysis spans various datasets, ranging from synthetic numerical simulations to those that closely mimic real-world scenarios.

3.4.1 Baseline angle selection Methods

This subsection introduces the foundational angle selection methods used as baselines in our X-ray CT studies:

- **Equidistant policy:** This method is a conventional strategy for determining the acquisition angles in X-ray CT imaging. It uniformly divides the full 360-degree rotation into equally spaced angular intervals. With an increase in the number of angles, the interval between consecutive angles becomes narrower, offering denser angular coverage.
- **Golden standard policy [49, 21]** In contrast to the equidistant approach, this method adopts a non-uniform strategy for angle selection, leveraging the golden ratio for angle distribution. This ensures that with the addition of angles, the previously established angles remain fixed, and new angles are allocated within the largest existing interval between angles, according to the golden ratio. This dynamic adjustment can optimize angular coverage.

3.4.2 Shepp-Logan phantoms

Initially, our experiment employs a numerical dataset and incorporates the CR metric into the reward function. The primary goal of this preliminary test is to demonstrate that our approach can effectively enhance defect detectability.

Dataset

For the first experiment, we generated images of objects consisting of ellipsoidal parts inspired by the Shepp-Logan phantom. The dataset consists of 1,800 images of size 128×128, evenly divided into 900 normal samples and 900 samples with elliptical defects. The shapes and defects are characterized by random variations in scale, shift,

and rotation. The Shepp-Logan shapes have scales within the following ranges: minor axis length between 57 and 67 pixels, and major axis length between 90 and 105 pixels. The defects have scales within these ranges: major axis length between 9 and 25 pixels, and minor axis length between 6 and 19 pixels. The rotations are distributed across 36 angles, each spaced equally from 0° to 179°. The diversity in the dataset is depicted in Figure 3.2.

Implementation

When computing the CR value, we consider an elliptical background area slightly larger than the elliptical defect, with both its major and minor axes extended by four pixels. For the stop criteria, we set the maximum number of angles, M, to 20, the threshold for reconstruction quality, Q, to 0.85, scale α, to 15 and penalty $-a$, to -1. The Astra-toolbox [90, 89] is employed for both forward projection and reconstruction tasks.

In the forward projection process, we utilize a fan-beam geometry. The distance from the source to the detector is set at 400 units, while the distance from the source to the object is 200 units. The detector is configured with a resolution of 256 pixels.

For the reconstruction phase, we employ SIRT with bound constraints, i.e., we run (3.3) with $\Lambda = \{\boldsymbol{x} \in \mathbb{R}^n | 0 \leq \boldsymbol{x}_i \leq 1\}$ for 150 iterations.

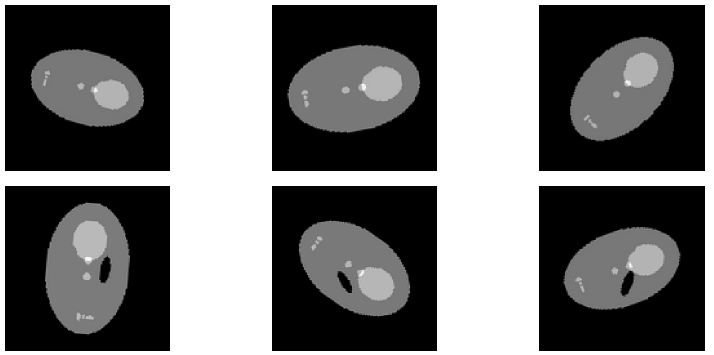


Figure 3.2: This figure displays samples of Shepp-Logan shapes, illustrating both normal (top row) and defective variants (bottom row). Each sample is unique, demonstrating a range of scales, shifts, and rotations, which reflects the diversity of the dataset used in our analysis.

Evaluation

During our training process, we conduct a series of 160,000 episodes to train the DRL agent. Each episode involves using a Shepp-Logan shape from our dataset. In Figure 3.3, the top demonstrates the progression of average total rewards accumulated by the DRL agent during the training phase. This trend indicates that the policy effectively learned to balance improving SSIM and contrast ratio values while reducing the number of angles over time.

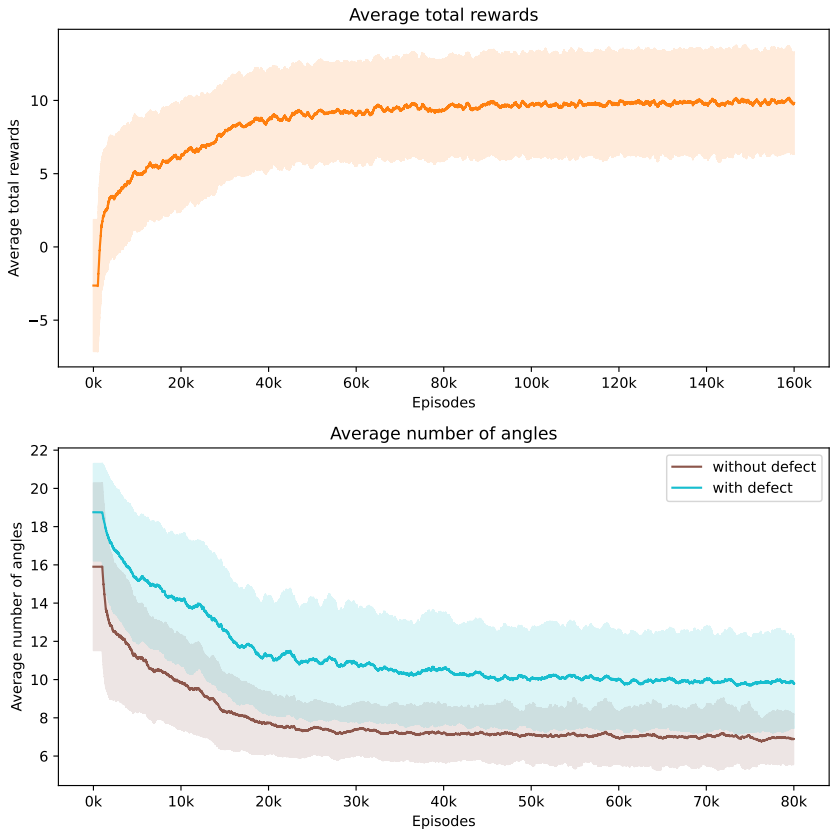


Figure 3.3: Evolution of average total rewards and the number of angles over training episodes. The figure at the top presents the trend in average total rewards achieved by the DRL agent throughout the training episodes. The figure at the bottom illustrates the variation in the average number of angles selected by the DRL agent for Shepp-Logan shapes, differentiated by the presence or absence of defects. Analyzed at intervals of every 1000 episodes, the graph shows mean values as a central curve, signifying the average reward at each point, with variance depicted as shaded bands around the curve, representing the standard deviation and the spread of reward values over time.

The bottom of the Figure 3.3 showcases how the number of angles chosen by the DRL agent adjusts in response to Shepp-Logan shapes, both with and without defects. Initially, the DRL agent tends to select a higher number of angles to meet the stop criteria. It is observed that the shapes with defects consistently require more angles compared to the normal shapes from the start of training. As training progresses, the number of angles decreases for both scenarios due to the negative reward applied at each step. When considered alongside Figure 3.3, this trend suggests that the DRL agent is optimizing its angle selection strategy by choosing fewer yet more informative angles, thereby enhancing the total rewards earned. We incorporate the contrast ratio into the reward function, which influences the agent's behavior when defects are present. In these cases, more angles are needed to enhance the visibility of flaws against the background. Notably, the data reveals that the defective Shepp-Logan shapes necessitate approximately 11 angles for effective analysis, compared to around 7 for non-defective ones.

Figure 3.4 presents a set of reconstructed image samples obtained using the DRL, equidistant, and golden standard policies. The number of angles used in each row is determined by the DRL policy. The first row, utilizing 8 angles, and the second row, employing 10 angles, highlight the superior clarity of defects in images reconstructed by the DRL policy.

The observed decrease in the number of angles chosen significantly influences the resulting CR values. Figure 3.5 contrasts the performance of three different policies —equidistant, golden standard, and DRL —in terms of their influence on CR value. The equidistant and golden standard policies utilize the number of angles determined by the DRL policy with SSIM and CR rewards. It is noted that while the average CR values under the equidistant and golden standard policies tend to diminish as the number of angles decreases, the CR values achieved using the DRL policy show a notable increase at the early stage and remain stable. Among these, the golden standard policy slightly outperforms the equidistant policy. Additionally, to evaluate the effectiveness of the CR reward, a separate DRL agent is trained solely with SSIM rewards. As shown in Figure 3.3, this approach initially leads to an increase in the CR values, which later declines. Upon reaching convergence, this SSIM-based DRL policy surpasses the equidistant and golden standard policies in performance but is still outperformed by the DRL policy, which incorporates a CR reward. Notably, the SSIM-based DRL policy focuses on a single task, leading to the selection of fewer angles than the DRL policy with dual tasks. Specifically, during the last 2000 episodes, the SSIM-based DRL policy selects an average of 8.07 angles, compared to 8.32 angles for

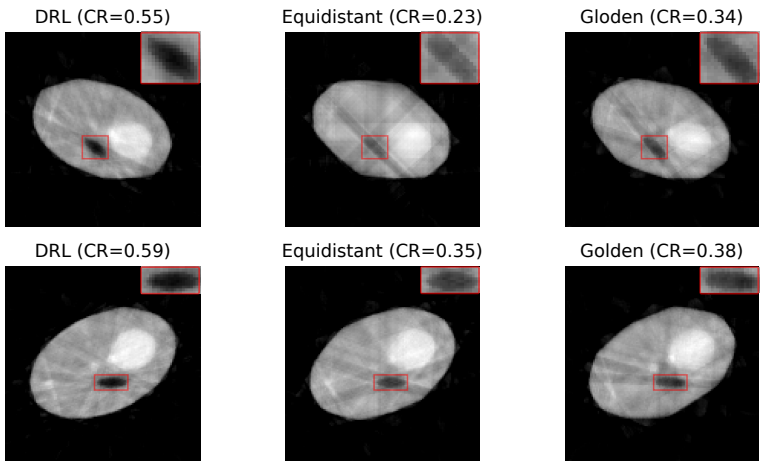


Figure 3.4: Reconstruction quality comparison. This figure presents a side-by-side comparison of reconstructed images obtained using the DRL policy versus those acquired through the equidistant and golden standard policies. The comparison highlights the differences in defect visibility across these methodologies. The numerical value accompanying each title denotes the CR value pertinent to its respective image, serving as a quantitative measure of the defect contrast compared to its surroundings. Red rectangles highlight the defect in each image, with a zoomed-in view displayed in the upper right corner to facilitate a more detailed examination of the defect visibility.

the DRL policy with the CR reward. The DRL policy can handle variations in scale and rotation, ensuring reconstruction quality and effective defect detection with an appropriate number of angles. These findings suggest that including a reward for CR in the DRL framework aids in adaptively selecting angles that are not only optimal in quantity but also more effective for defect detection.

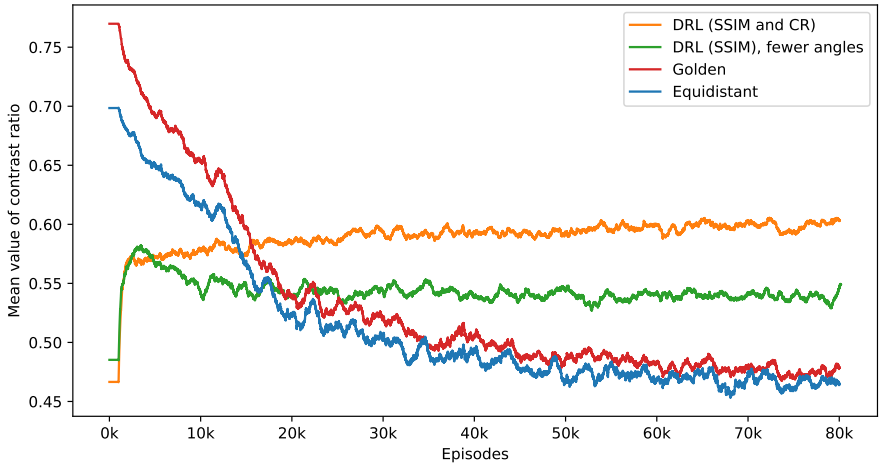


Figure 3.5: This figure presents a side-by-side comparison of CR values derived from different imaging policies. It specifically highlights the differences in CR values achieved by the DRL policy that utilizes both SSIM and CR as rewards, in contrast to the equidistant and golden standard policies. The number of angles selected by the DRL policy, which integrates SSIM and CR rewards, forms the basis for this comparison. The DRL policy informed by SSIM alone selects a smaller number of angles. It takes into account 1000 episodes, with the mean values representing the average. Additionally, calculations not depicted in this figure indicate that the DRL policy exhibits the smallest variance among these policies once the mean values reach a point of convergence.

To assess the efficacy of the trained policy, we examine the same phantom subjected to previously unseen rotations. According to Table 3.1, the DRL policy with dual tasks demonstrates superior performance compared to two traditional approaches, across all policies considering the average number of angles, which is 8.46 as determined by the DRL policy. Additionally, both the golden ratio and equidistant policies require the selection of three additional angles to achieve performance comparable to that of the DRL policy with dual tasks, as indicates in the last two rows of Table 3.1.

Table 3.1: Comparative performance of DRL policy with dual tasks and traditional angle selection policies on Shepp-Logan phantom.

Policy	Average number of angles	CR	SSIM
DRL	8.46	0.62 ± 0.10	0.86 ± 0.01
Golden ratio	8.46	0.48 ± 0.13	0.83 ± 0.01
Equidistant	8.46	0.49 ± 0.17	0.83 ± 0.01
Golden ratio	8.46+3	0.61 ± 0.11	0.85 ± 0.01
Equidistant	8.46+3	0.59 ± 0.13	0.86 ± 0.01

3.4.3 Simulated industrial dataset

After the proof of the first experiment focusing on enhancing the contrast of the defect, a second experiment is carried out to assess the pipeline against a more realistic scenario and involve the defect segmentation.

The Fraunhofer EZRT XSimulation software [68] is used to acquire a set of 1800 simulated images of size 128×128 in fan beam geometry. These simulated images serve as the ground truth and are used to compare with the K-means clustering segmentation to calculate the Dice score values. In this study, we focus on two prevalent defect types —pores and cracks—to evaluate the efficacy of the DRL approach. Within the domain of industrial CT scans, Regions of Interest (ROIs) are typically predetermined. Certain areas are more prone to defects and, consequently, demand a higher reconstruction quality. To reflect this industry practice, we designate two distinct ROIs to separately introduce pore and crack defects. Mirroring the setting of our initial experiment, we consider an equal distribution of normal and defective samples, totaling 900 each. Figures 3.6 and 3.7 display examples of these artificially induced defects for reference. Variable position and tilt of the object are chosen for data augmentation and bias reduction.

In this research, we introduce pore defects of varying sizes into ROI 1. These defects encompass a pixel count between 11 and 35, with circumcircle diameters ranging from 3.2 to 7.0 pixels. This range presents diverse scenarios for the DRL method to tackle. Similarly, cracks are inserted into ROI 2, with a pixel count ranging from 7 to 25 and circumcircle diameters between 8.6 and 19.7 pixels. These experiments are designed to rigorously assess the DRL method's effectiveness in segmenting defects and enhancing the quality of reconstructions within ROIs.

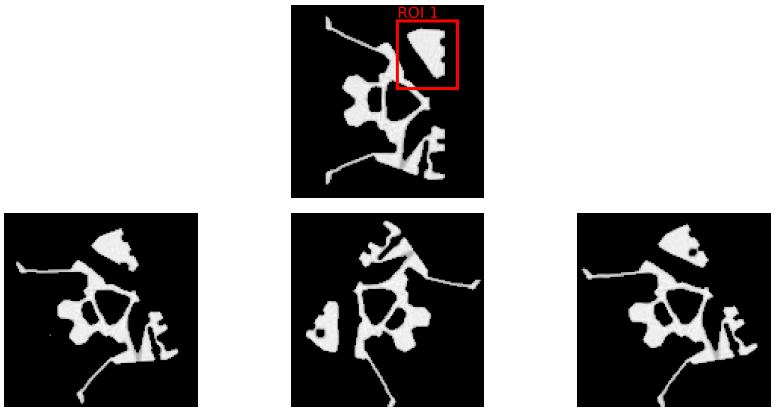


Figure 3.6: The top row depicts the selected ROI from the initial dataset, devoid of any defects. The bottom row presents three samples with artificially inserted pore defects within the ROI, each sample exhibiting a unique combination of rotation and scale variations to simulate defect diversity.

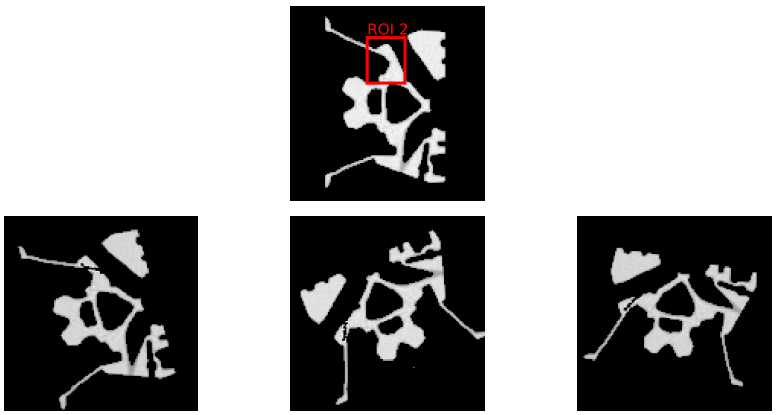


Figure 3.7: The top row depicts the selected ROI from the initial dataset, devoid of any defects. The bottom row presents three samples with artificially inserted crack defects within the ROI, each sample exhibiting a unique combination of rotation and scale variations to simulate defect diversity.

Implementation

In this investigation, we assess the utility of our approach in augmenting the defect segmentation capabilities of the K-means clustering algorithm. Because the sample investigated in Section 4.3 has homogeneous materials, the number of clusters is two. Additionally, We identify the pixel with the maximum intensity value and use its position as the initial center for one of the clusters. This approach helps distinguish between the defect and non-defect areas. Finally, we compare the segmented results with ground truth annotations to validate the clustering performance using the DS metric.

To establish clear termination parameters for our process, we define two stopping criteria: a maximum count of 20 angles (M) and a minimum threshold for reconstruction quality within the ROIs, set at a SSIM of 0.45 (Q). To ensure uniformity throughout our experiments, we adhere to the same configuration settings used in our initial experiment, both in other parameters and within the Astra Toolbox. Additionally, we train two distinct DRL policies to address pore and crack defects, respectively, allowing for a targeted approach to each defect type and ROI.

Evaluation

A. Pore defects

Our investigation centers on pore defects within ROI 1. Figure 3.8 captures the evolution of the training phase, tracking both the average total rewards and the number of angles utilized. The upper section of Figure 3.8 demonstrates that the DRL agent successfully progressively increases its reward acquisition. The lower section reveals the strategy of the agent to minimize the number of angles used. Notably, non-defective samples tend to require a greater number of angles on average compared to defective ones, and they also exhibit a broader variance.

Figure 3.9 presents two sets of reconstructed images using the DRL policy, alongside those derived from the equidistant and golden standard policies. This comparison is based on the number of angles determined by the DRL policy. The first row, utilizing 10 angles, and the second row, employing 7 angles, are evident that the reconstructions produced by the DRL policy are superior, demonstrating enhanced defect visualization. This improvement is corroborated by the corresponding DS values, indicating a higher accuracy of defect detection.

Subsequently, Figure 3.10 showcases the segmentation outcomes corresponding to the top group of reconstructions from Figure 3.9. The reconstructed images are pro-

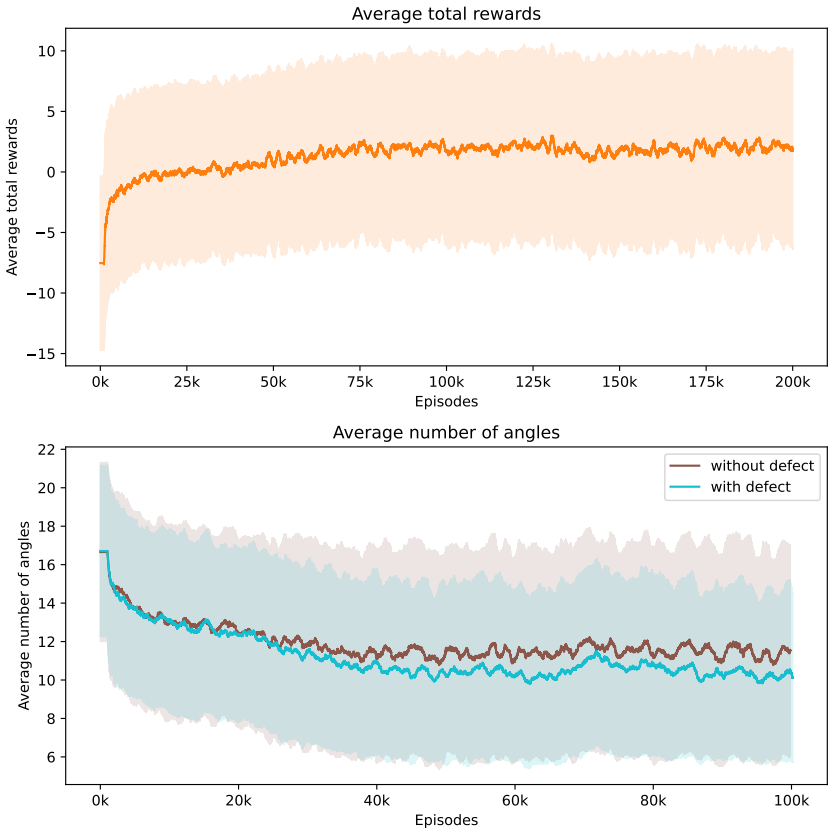


Figure 3.8: Evolution of average total rewards and the number of angles over training episodes. The figure at the top presents the trend in average total rewards achieved by the DRL agent throughout the training episodes. The figure at the bottom illustrates the variation in the average number of angles selected by the DRL agent for samples with ROI 1, differentiated by the presence or absence of defects. Analyzed at intervals of every 1000 episodes, the graph shows mean values as a central curve, signifying the average reward at each point, with variance depicted as shaded bands around the curve, representing the standard deviation and the spread of reward values over time.

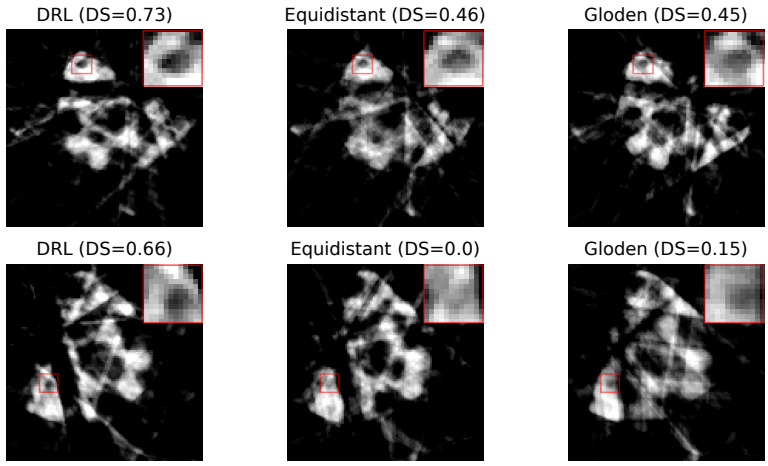


Figure 3.9: Reconstruction quality comparison. This figure presents a side-by-side comparison of reconstructed images obtained using the DRL policy versus those acquired through the equidistant and golden standard policies. The comparison highlights the differences in defect visibility across these methodologies. The numerical value accompanying each title denotes the DS value pertinent to its respective image, serving as a quantitative measure of the defect segmentation. Red rectangles highlight the defect in each image, with a zoomed-in view displayed in the upper right corner to facilitate a more detailed examination of the defect visibility.

cessed via the K-means clustering algorithm, resulting in binary representations of the defects. Segmentation accuracy is then ascertained by comparing these binary images against the ground truth defect masks, demonstrating the practical effectiveness of our DRL policy in identifying pore defects within the examined ROI.

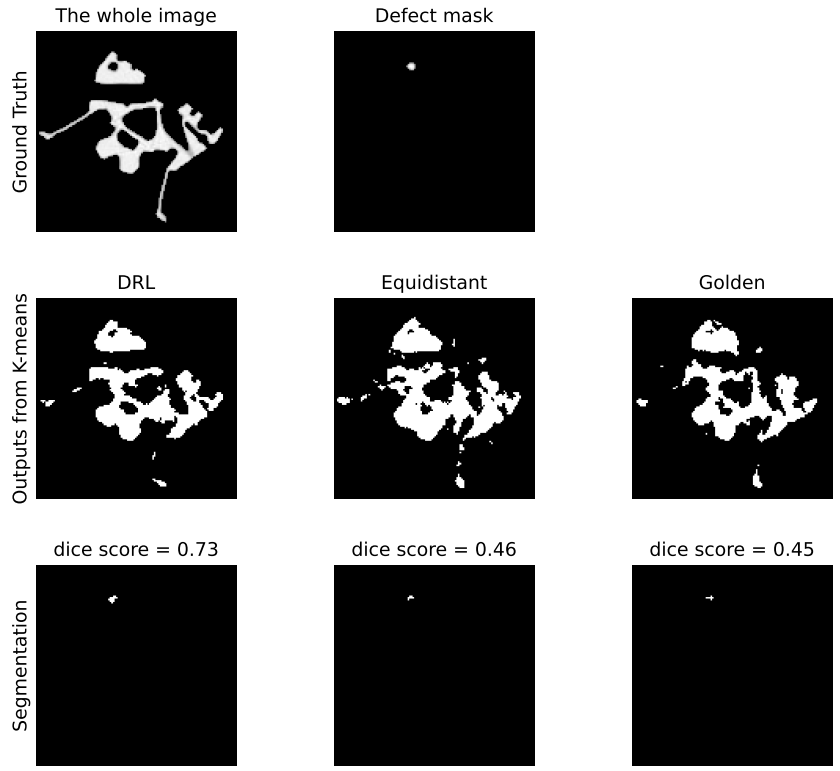


Figure 3.10: Comparative analysis of the pore defect segmentation. The top row represents the original image with ground truth defects. The middle row illustrates segmentation outputs using three different policies: DRL policy, equidistant policy, and a golden standard policy. The bottom row displays the corresponding defect masks generated by K-means clustering. The values of the DS for each method are indicated, quantifying the accuracy of the defect segmentation relative to the ground truth.

Figure 3.11 presents the impact of reducing the number of selected angles on the DS value. This figure provides a performance comparison of three distinct policies—equidistant, golden standard, and DRL. The number of angles used by these policies is set according to the DRL policy, which integrates both SSIM and DS rewards. It is observed that while the DS values for the equidistant and golden standard policies generally decrease with a reduction in the number of angles, the DRL policy either

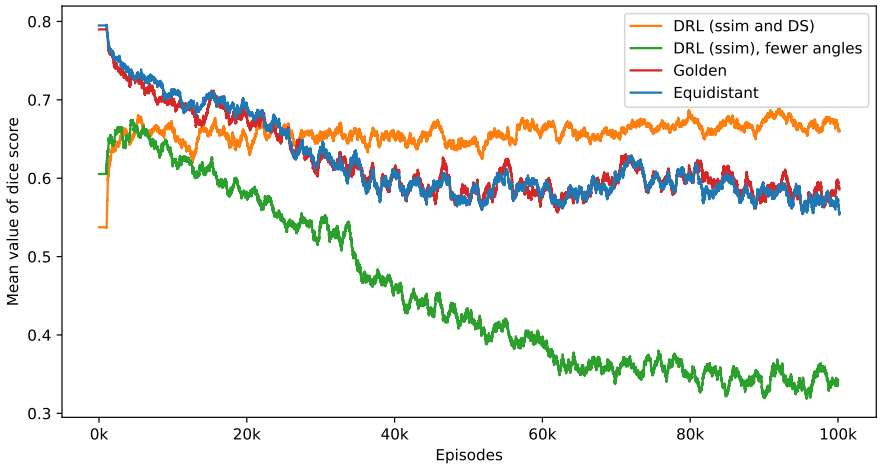


Figure 3.11: This figure presents a side-by-side comparison of DS values derived from different imaging policies. It specifically highlights the differences in DS values achieved by the DRL policy that utilizes both SSIM and DS as rewards, in contrast to the equidistant and golden standard policies. The number of angles selected by the DRL policy, which integrates SSIM and DS rewards, forms the basis for this comparison. The DRL policy informed by SSIM alone selects a smaller number of angles. It takes into account 1000 episodes, with the mean values representing the average. Additionally, calculations not depicted in this figure indicate that the DRL policy exhibits the smallest variance among these policies once the mean values reach a point of convergence.

sustains or enhances its DS values. Notably, the golden standard policy demonstrates a performance level that is similar to that of the equidistant policy. The effectiveness of the DRL policy, integrating both SSIM and DS rewards, is clearly demonstrated. In contrast, a DRL agent trained exclusively with the SSIM as its reward displays inconsistent performance, initially showing an increase in DS values but subsequently declining. When this training process reaches its endpoint, the SSIM-based DRL policy is outperformed by both equidistant and golden policies. As mentioned in the first experiment, the SSIM-based DRL policy necessitates fewer angles, averaging 5.65 over the final 2000 episodes, in contrast to the DRL policy that integrates both SSIM and DS rewards, which averages 10.84 angles. This difference in the number of angles is a contributing factor to the lesser performance observed in the SSIM-based policy. These findings validate the importance of including a DS reward in the DRL algorithm, significantly enhancing the selection of angles for improved defect segmentation. Without this reward, the DRL approach tends to focus solely on the reconstruction quality within the ROI while overlooking the specifics of the defect.

In a manner analogous to the experiments conducted on the Shepp-Logan phantom, we evaluate the trained DRL policy with dual tasks using rotations that are not included in the training set. The DRL policy selects an average of 9.84 angles. Table 3.2 shows that this policy outperforms two traditional policies in terms of performance. Furthermore, the last two rows of Table 3.2 indicate that both the golden ratio policy and the equidistant policy require the selection of two additional angles to achieve performance comparable to that of the DRL policy with dual tasks.

Table 3.2: Comparative performance of DRL policy and traditional angle selection policies on the simulated industrial dataset with pore defects.

Policy	Average number of angles	CR	SSIM
DRL (9.84)	9.84	0.65 ± 0.24	0.46 ± 0.03
Golden ratio	9.84	0.57 ± 0.32	0.43 ± 0.09
Equidistant	9.84	0.56 ± 0.34	0.43 ± 0.07
Golden ratio	9.84+2	0.67 ± 0.25	0.47 ± 0.07
Equidistant	9.84+2	0.67 ± 0.28	0.47 ± 0.06

B. Crack defects

Crack defects, with their distinctively long and thin shapes, pose unique challenges for automated detection and segmentation systems. In an experiment analogous to that conducted for pore defects, we observe that the average total rewards for the crack

defect samples increase over time while the average number of angles used decreases, as shown in Figure 3.12. Interestingly, the number of angles required for accurately detecting defective samples is similar to that for non-defective samples in the case of crack defects.

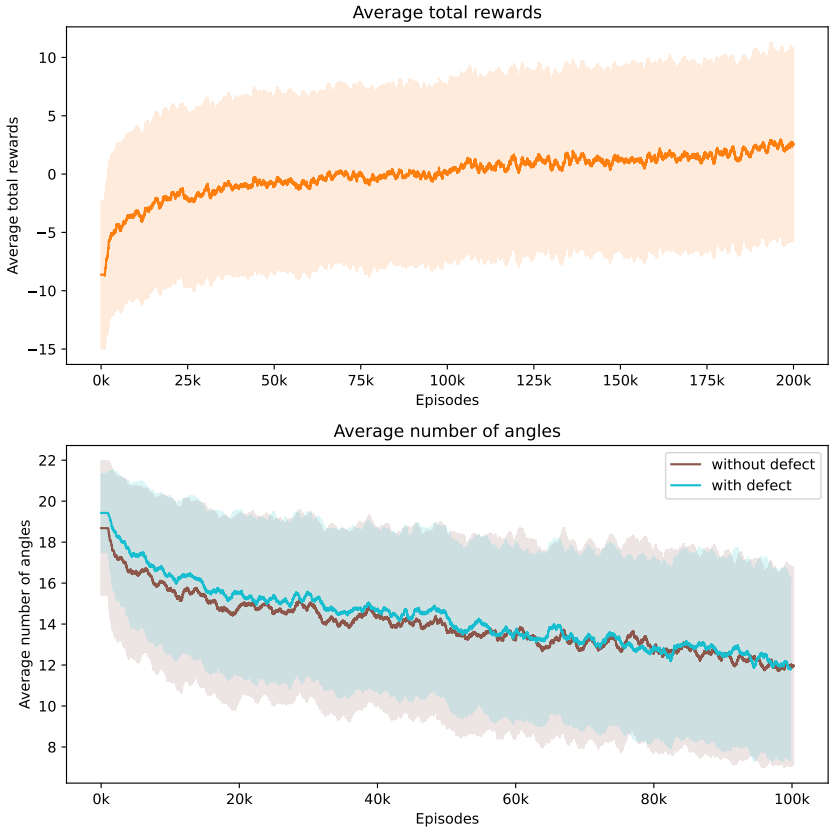


Figure 3.12: Evolution of average total rewards and the number of angles over training episodes. The figure at the top presents the trend in average total rewards achieved by the DRL agent throughout the training episodes. The figure at the bottom illustrates the variation in the average number of angles selected by the DRL agent for samples with ROI 2, differentiated by the presence or absence of defects. Analyzed at intervals of every 1000 episodes, the graph shows mean values as a central curve, signifying the average reward at each point, with variance depicted as shaded bands around the curve, representing the standard deviation and the spread of reward values over time.

In our analysis, Figure 3.13 displays the reconstructed images of crack defects using three distinct policies: the DRL policy, and the equidistant and golden standard policies. Additionally, the number of angles is determined by the DRL policy. The first

row utilizes 14 angles, while the second row employs 18 angles. Similar to pore defects, the DRL policy, as indicated by the visibility of the defects and their corresponding DS values, appears to significantly enhance the segmentation process, making cracks more distinguishable for subsequent analysis.

Figure 3.14 presents the outcomes of applying the K-means clustering algorithm for segmenting these crack defects. The outcomes are corresponding to the first row in Figure 3.13. As evidenced by the DS values, the segmentation performance illustrates the DRL policy's superior ability to accurately define and isolate the cracks within the material sample, underscoring its potential in improving segmentation in CT imaging. Additionally, Figure 3.15 presents a similar impact of reducing the number of selected angles on the DS value to Figure 3.11. When examining the last 2000 episodes, the average number of angles determined by the DRL policy, which incorporates both SSIM and DS rewards, is 11.86. In contrast, the average for the DRL policy that relies solely on SSIM is 9.39. However, it is noteworthy that the performance of the SSIM-based DRL policy exhibits a degree of instability.

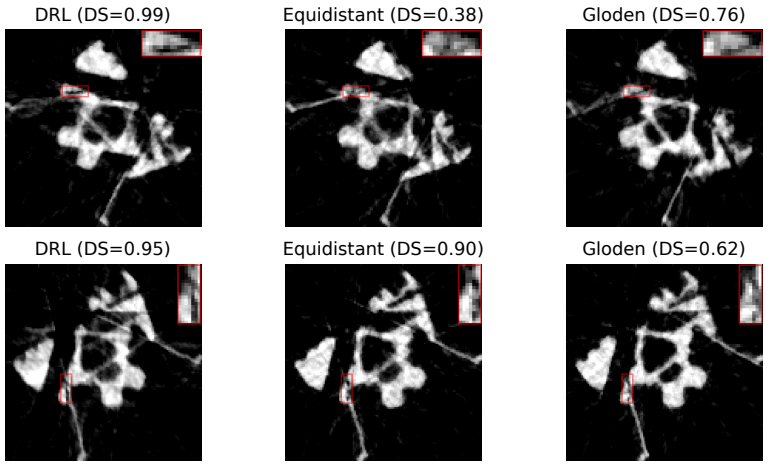


Figure 3.13: Reconstruction quality comparison. This figure presents a side-by-side comparison of reconstructed images obtained using the DRL policy versus those acquired through the equidistant and golden policies. The comparison highlights the differences in defect visibility across these methodologies. The numerical value accompanying each title denotes the DS value pertinent to its respective image, serving as a quantitative measure of the defect segmentation. Red rectangles highlight the defect in each image, with a zoomed-in view displayed in the upper right corner to facilitate a more detailed examination of the defect visibility.

In the context of a simulated industrial dataset featuring crack defects, the DRL

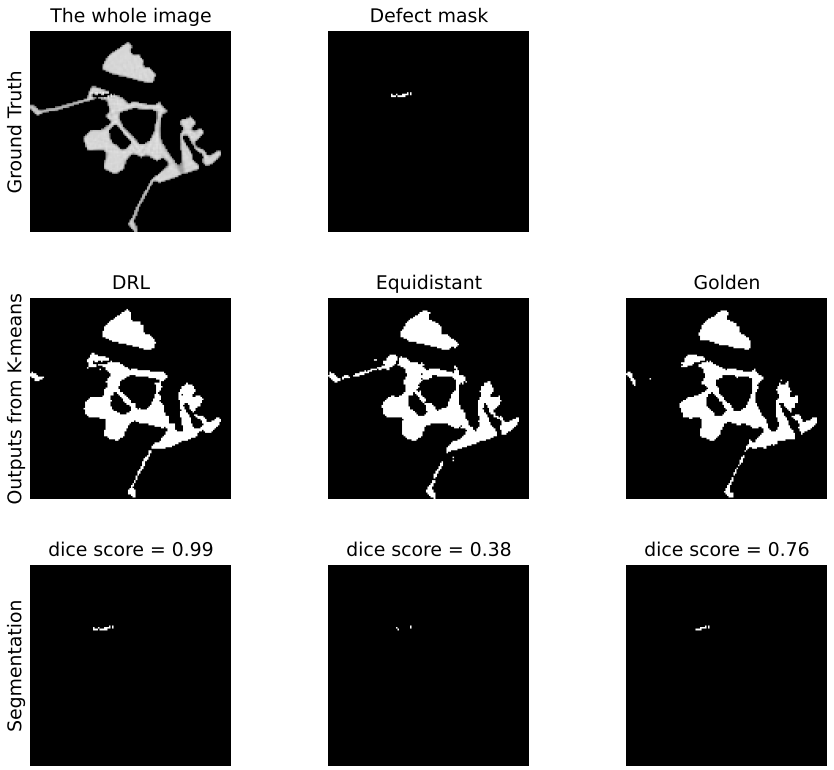


Figure 3.14: Comparative analysis of the crack defect segmentation. The top row represents the original image with ground truth defects. The middle row illustrates segmentation outputs using three different policies: DRL policy, equidistant policy, and a golden standard policy. The bottom row displays the corresponding defect masks generated by K-means clustering. The values of the dice score for each method are indicated, quantifying the accuracy of the defect segmentation relative to the ground truth.

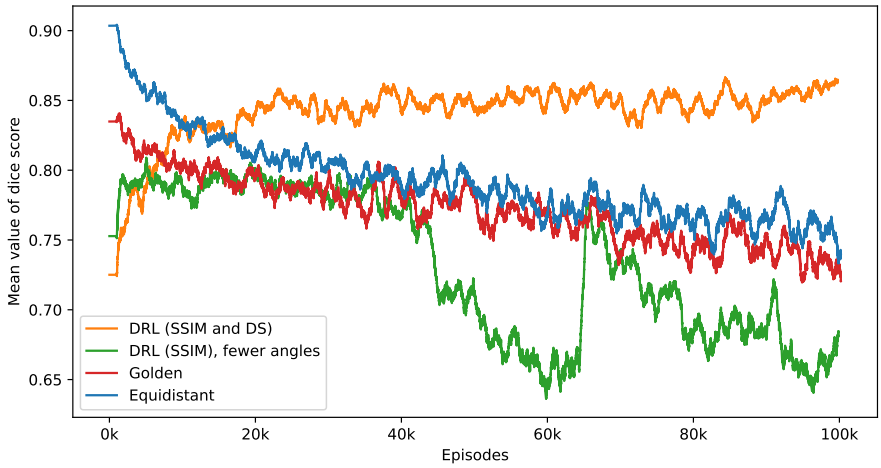


Figure 3.15: This figure presents a side-by-side comparison of DS values derived from different imaging policies. It specifically highlights the differences in DS values achieved by the DRL policy that utilizes both SSIM and DS as rewards, in contrast to the equidistant and golden standard policies. The number of angles selected by the DRL policy, which integrates SSIM and DS rewards, forms the basis for this comparison. The DRL policy informed by SSIM alone selects a smaller number of angles. It takes into account 1000 episodes, with the mean values representing the average. Additionally, calculations not depicted in this figure indicate that the DRL policy exhibits the smallest variance among these policies once the mean values reach a point of convergence.

policy with dual tasks selects an average of 12.26 angles. As indicated by Table 3.3, this policy surpasses the performance of two traditional policies. Additionally, the last two rows of Table 3.3 reveal that to match the performance of the DRL policy with dual tasks, both the golden ratio policy and the equidistant policy necessitate the selection of six additional angles.

Table 3.3: Comparative performance of DRL policy with dual tasks and traditional angle selection policies on the simulated industrial dataset with crack defects.

Policy	Average number of angles	CR	SSIM
DRL	12.26	0.85 ± 0.14	0.46 ± 0.03
Golden ratio	12.26	0.72 ± 0.28	0.39 ± 0.10
Equidistant	12.26	0.76 ± 0.25	0.40 ± 0.10
Golden ratio	12.26+6	0.83 ± 0.18	0.49 ± 0.07
Equidistant	12.26+6	0.86 ± 0.14	0.49 ± 0.07

3.5 Discussion

Our study initiates with numerical experiments using the Shepp-Logan phantom, featuring ellipse-shaped defects, and employs a CR-based reward to initially evaluate how defect information influences the DRL policy. Advancing towards more complex scenarios, we tackle the challenge of detecting smaller defects, which are inherently more difficult to identify. Our investigation encompasses two predominant defect types—pores and cracks—and incorporates K-means clustering for the segmentation of these defects. Following this, we integrate the DS metric into the reward function, showcasing its effectiveness in boosting the K-means clustering algorithm's ability to segment defects. The findings reveal that DRL can successfully learn task-specific sequential OED policies that surpass traditional equidistant and golden standard approaches. This achievement underscores the potential to include additional tasks in our methodology, allowing for more tailored angle selection adaptations. In such cases, carefully optimizing the assigned weights for these tasks is crucial for achieving balanced and effective results.

Another notable aspect of our approach is the introduction of a negative reward at each step, encouraging the DRL agent to select an optimal number of angles. This strategy balances experimental cost and quality, proving advantageous for practical applications. The experimental results consistently surpass the outcomes of traditional

equidistant and golden standard methods, underscoring the DRL approach's ability to effectively select a flexible number of angles for improved defect detection.

To apply our method to real-world settings, the DRL approach has to be extended to incorporate one more component: To focus on task adaptation, a simple stopping criterion was used in this work, which depended on the ground truth images. In real-world settings, the DRL policy also needs to decide when to stop. One of the approaches for this that we will investigate in the future is to use the value function of the DRL framework, which is an estimate of future total rewards, as a stopping criterion, e.g., by setting a tolerance level for the output of the value function [85].

Several challenges and future works need to be addressed to advance this method. CAD models would be valuable for offline training of the DRL agent. However, it is essential to build a reliable simulation environment using CAD models that closely mimics real-world conditions. Besides the metrics involved in the reward function, such as those assessing reconstruction quality and defect detectability, other metrics for industrial quality control can also be considered. Then, we aim to extend our method to encompass 3D geometries, in particular the cone-beam geometry, which is most commonly used in industrial applications. We also plan to investigate more dynamic scanning trajectories, including out-of-plane angles and variable focus, to align with the advanced capabilities of robotic CT scanners in industrial inspections. These developments promise to increase the versatility and applicability of our approach across diverse industrial inspection scenarios. Another aspect worth considering is the impact of the reconstruction algorithm on the efficiency of inline CT scans. Exploring deep-learning-based reconstruction methods [6] instead of traditional iterative methods like SIRT could potentially speed up the process, leading to more efficient industrial CT applications. Additionally, more efficient training methods, particularly reducing the computational complexity for handling large-scale data, need to be considered. Finally, adopting more advanced defect segmentation methods could improve our ability to manage the complexities associated with diverse defect types and imaging conditions.

3.6 Conclusion

This paper aims to expand the complexity of the reward function used in DRL for sequential OED and validate its effectiveness through various experiments. The inclusion of a reward for defect information enables the DRL agent to effectively learn task-specific angle selection. Moreover, our method offers flexibility in the number of angles utilized, achieving comparable reconstruction quality and defect detectability

with fewer angles than those required by golden ratio and equidistant policies. This methodology shows promise for online application in practice following the training phase.

In conclusion, our research sets the stage for future advancements in applying DRL to various industrial CT scanning scenarios, with the potential to improve the accuracy and efficiency of automated inspection technologies.

Chapter 4

Dynamic Angle Selection in X-Ray CT: A RL Approach to Optimal Stopping

4.1 Introduction

X-ray Computed Tomography (CT) enables inline industrial inspection through three-dimensional reconstruction. However, fast and adaptive CT scanning is essential to make its widespread industrial application feasible. Previous studies have shown that not all projections are equally informative for certain objects [46, 91]. Most research has focused on sequentially selecting informative angles to enhance efficiency. This sequential Optimal Experimental Design (sOED) is often described within a Bayesian framework [53, 19, 64], where angles are chosen to maximize information gain. Information gain is typically quantified by comparing the prior and posterior distributions of the reconstruction or by assessing the similarity between the reconstructed image and the ground truth.

This chapter is based on:

Tianyuan Wang, Felix Lucka, Daniël M. Pelt, K. Joost Batenburg, and Tristan van Leeuwen. "Dynamic Angle Selection in X-Ray CT: A Reinforcement Learning Approach to Optimal Stopping". In: *Applied Mathematics for Modern Challenges* 5 (2025), pp. 36–63. DOI: 10.3934/ammc.2025010

Batenburg *et al.* [9] and Dabravolski *et al.* [22] used a set of template images composed of Gaussian blobs to represent samples from the prior distribution. They introduced an upper bound [8] to approximate the information gain, which reflects the diameter of the solution set. Burger *et al.* [14] employed classical "alphabetic criteria" for OED, such as "A-" and "D-" optimality, using the trace or determinant of the covariance matrix of the posterior distribution as summary statistics [43]. Additionally, they used a Gaussian distribution as a prior, updating the posterior after selecting each angle. Building on this approach, Helin *et al.* [39] introduced a Total Variation (TV) prior to enhance edges in reconstructions. The non-Gaussian TV prior was approximated as a Gaussian distribution using lagged diffusivity iteration. Furthermore, Barbano *et al.* [7] utilized a deep image prior as the reconstruction method and linearized the network to approximate the posterior distribution as a Gaussian. To avoid the need for a closed-form solution, Elata *et al.* [26] proposed using a diffusion model for CT reconstruction, sampling from the posterior to approximate the posterior covariance matrix.

Recently, policy-based methods from the reinforcement learning community have been introduced into sOED [77, 13, 30]. In the medical CT field, Shen *et al.* [78] trained a gated recurrent unit as a policy network on simulated medical data to map projections to probabilities over the angle space. In previous work, we explored industrial CT applications with very few angles [97]. We trained a policy that maps the current reconstruction to probabilities over the angle space. Additionally, we addressed a specific task—defect detection—by incorporating an extra reward for defect detectability and prior information about defects, enabling the trained policy to identify informative angles to aid in defect detection [96].

Although extensive work has been devoted to selecting informative projection angles, the question of how many angles to acquire is often neglected. This choice can be cast as an optimal stopping problem, a topic well studied in financial mathematics. The optimal expected reward is given by the Snell envelope, the smallest super-martingale that dominates the reward process. The earliest (respectively, latest) optimal stopping time is the first instant at which the immediate reward equals (respectively, exceeds) the continuation value [63, 11, 23].

Additionally, most studies have focused on simulated data rather than experimental X-ray CT data. This is particularly evident in learning-based methods, which rely on training with simulated data, leaving their generalizability to experimental X-ray CT data uncertain.

The contributions of this work include the development of an optimal stopping

method to balance the trade-off between experimental costs and experimental goals, such as reconstruction quality. This approach enables both adaptive selection of informative angles and optimal scan termination based on experimental costs. By incorporating a terminal policy into the Actor–Critic framework, we proposed a novel method for computing the policy gradient, jointly optimizing angle selection and the terminal policy. Additionally, we investigated the gap between simulation and real-world applications by evaluating the trained model on experimental X-ray CT data.

The structure of the paper is as follows: The Background section provides an overview of the fundamental concepts and notations related to inverse problems and sOED using reinforcement learning. The Method section introduces our novel approach for computing the policy gradient. The Results section presents the findings from both simulation and experimental X-ray CT data experiments. Finally, the paper concludes with a discussion of the key findings and their implications.

4.2 Background

4.2.1 Forward and Inverse Problems and Optimal Experimental Design (OED)

Measurements $\boldsymbol{y}(\boldsymbol{\theta})$ are obtained from the ground-truth image (underlying parameters) $\bar{\boldsymbol{x}}$ using a forward operator $\boldsymbol{A}(\boldsymbol{\theta})$, which is determined by the design parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$. In the case of X-ray CT, the forward operator $\boldsymbol{A}(\boldsymbol{\theta})$ corresponds to the Radon transform for angles $\boldsymbol{\theta}$ [38]. Since the measurement process is subject to noise, we also incorporate a noise term $\boldsymbol{\epsilon}(\boldsymbol{\theta})$ and model the projections as follows:

$$\boldsymbol{y}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\bar{\boldsymbol{x}} + \boldsymbol{\epsilon}(\boldsymbol{\theta}). \tag{4.1}$$

The inverse problem involves using the projections $\boldsymbol{y}(\boldsymbol{\theta})$ and the forward operator $\boldsymbol{A}(\boldsymbol{\theta})$ to compute a reconstruction $\widehat{\boldsymbol{x}}(\boldsymbol{\theta})$ as an estimate of the ground-truth image $\bar{\boldsymbol{x}}$. Figure (4.1, a) illustrates the forward and inverse processes. However, solving inverse problems is often challenging due to their ill-posedness [58]. The accuracy of the reconstructed ground-truth image is influenced by the design parameters $\boldsymbol{\theta}$, making their optimal selection crucial. OED is employed to select the most informative angles to acquire the corresponding projections. To quantify the accuracy of the reconstruction, a utility function is defined within the OED framework. The optimal design $\boldsymbol{\theta}^*$ is then obtained by maximizing the expected value of this utility function over the design space, taking into account the projections $\boldsymbol{y}(\boldsymbol{\theta})$ and the ground truth image $\bar{\boldsymbol{x}}$ [53,

67]. As shown in Figure (4.1, b), we plot reconstruction quality—measured by PSNR —versus the number of projection angles M (up to 50). Reconstructions $\widehat{\boldsymbol{x}}(\boldsymbol{\theta})$ are computed with the Simultaneous Iterative Reconstruction Technique (SIRT) under a nonnegativity constraint (150 iterations). The orange curve corresponds to a selection that, at each step, exhaustively evaluates all remaining angles and chooses the most informative one; the blue curve uses uniformly spaced angles. For the triangle phantom (left), the selection consistently outperforms the uniform strategy, especially at small M. The uniform strategy occasionally picks informative directions, which explains the oscillations in its performance. For the Shepp–Logan phantom (right), which is approximately rotationally symmetric and lacks a preferred direction, the gap between the two strategies is small. Across both phantoms we observe diminishing returns: beyond a moderate number of angles, further increases in M yield only marginal PSNR gains. When acquisition is costly, such small gains may not justify additional projections, motivating an optimal stopping rule to balance reconstruction quality against experimental cost.

4.2.2 Sequential Optimal Experimental Design (sOED) and Reinforcement Learning

Traditional experimental design is typically performed a-priori, with all optimal design parameters selected simultaneously. Because it ignores the feedback obtained after each parameter choice, this approach cannot support a-posteriori selection, potentially overlooking information that could refine subsequent decisions [77].

The sOED extends the concept of traditional OED by allowing the design parameters to be determined sequentially, based on the data acquired from previous projections [19, 67]. For example, the design parameter θ_k can be chosen based on the previous projections $\boldsymbol{y}(\{\theta_1, \ldots, \theta_{k-1}\})$. This approach enables the design process to adapt dynamically to changes in the underlying parameters (ground-truth image) by adjusting the design parameters iteratively.

Solving OED problems in image reconstruction is inherently a bi-level optimization challenge: the lower level reconstructs the image, while the upper level minimizes the reconstruction error relative to the ground truth. The joint optimization problem is characterized by nonconvexity, nonlinearity, and high dimensionality, particularly in imaging applications [66]. The sequential approach further complicates this optimization, as the bi-level optimization problem would need to be solved in real time. Reinforcement Learning (RL) [84], a machine learning technique designed for fast se-

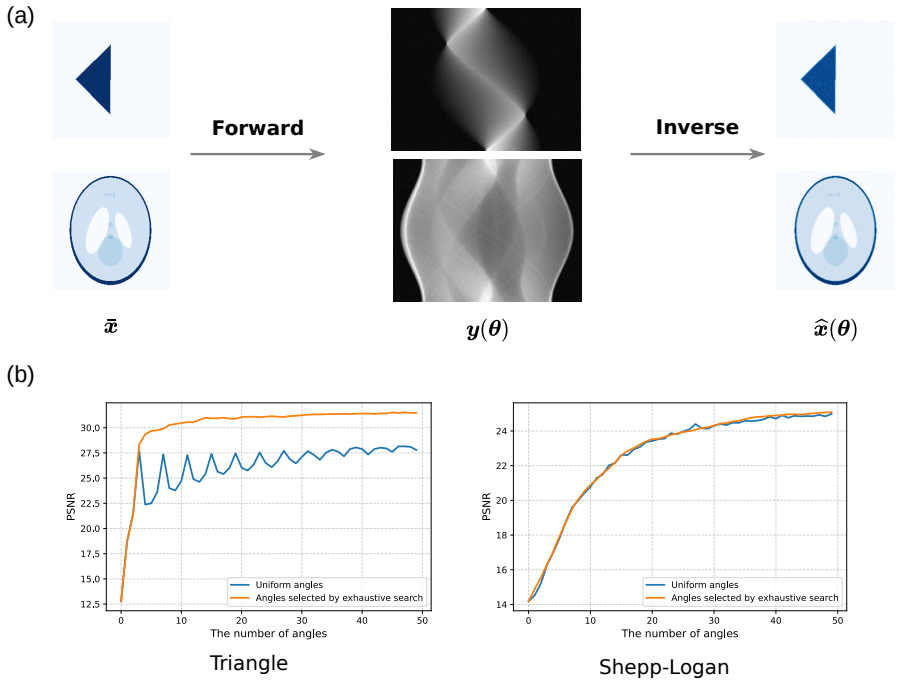


Figure 4.1: Triangle phantom example. (a) Forward process: Noisy parallel-beam projections at 180 equally spaced angles; additive zero-mean Gaussian noise (5% level) is applied to the noise-free sinogram. Inverse process: the image is reconstructed from these noisy projections. (b) PSNR as a function of the number of angles for the triangle phantom. The number of angles increases in increments of one. The orange curve represents angles selected via exhaustive search, while the blue curve corresponds to uniformly spaced angles. Left: triangle phantom; right: Shepp–Logan phantom.

quential decision-making, facilitates the resolution of this sOED by training through interaction with the environment. RL is grounded in the framework of Markov Decision Processes (MDPs), which consist of a state space, action space, transition model, and reward function. In practice, MDPs are sometimes extended to Partially Observable MDPs (POMDPs) when the underlying parameters are not fully observable, necessitating the reconstruction of the belief state from measurements. The goal of RL is to learn a parameterized policy that maps the current state to the next action, maximizing the expectation of the cumulative rewards [84].

Figure (4.2) illustrates how the RL framework integrates with the sOED process. In the context of sOED, at the k^{th} step, the action space consists of the possible values that the design parameter θ_k can take. After selecting θ_k, the reconstructed underlying parameters $\widehat{\boldsymbol{x}}_{k+1}$, inferred from all previous projections $\boldsymbol{y}(\{\theta_1, \ldots, \theta_k\})$, serve as the belief state. The utility function used for accuracy estimation acts as the reward function R. For example, R can be defined as PSNR, which estimates the reconstruction quality by comparing the reconstruction with the ground-truth image [97]. Consequently, the optimization problem of sOED is reformulated as learning a parameterized policy $\pi_{\text{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$, where $\boldsymbol{w}_a$ represents the policy parameters. This policy, with the optimal parameters $\boldsymbol{w}_a^*$, maximizes the expectation of the cumulative rewards obtained from the experiments.

By employing RL, we optimize *policy parameters* rather than the design parameters directly. We consider a finite-horizon return $\sum_{k=1}^{M} r_k$. The state-value function $V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_1)$ is the expected cumulative reward from the initial state $\widehat{\boldsymbol{x}}_1$ under the trajectory distribution π_{chain}, which factorizes into the action policy and the data-driven transition:

$$V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_1) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_{\text{chain}}} \left[\sum_{k=1}^{M} r_k \mid \widehat{\boldsymbol{x}}_1 \right], \tag{4.2}$$

where the trajectory distribution is

$$\pi_{\text{chain}}(\boldsymbol{\tau}; \boldsymbol{w}_a) = \prod_{k=1}^{M} \pi_{\text{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \ \cdot \ \pi_t(\widehat{\boldsymbol{x}}_{k+1} \mid \widehat{\boldsymbol{x}}_k, \theta_k), \tag{4.3}$$

and π_t denotes the acquisition+reconstruction state transition (unknown analytically; we sample it by taking a projection at θ_k and running one reconstruction update).

A trajectory $\boldsymbol{\tau}$ of M steps, $\{\widehat{\boldsymbol{x}}_1, (\theta_1, \widehat{\boldsymbol{x}}_2, r_1), \ldots, (\theta_M, \widehat{\boldsymbol{x}}_{M+1}, r_M)\}$, is thus generated by iterating: choose $\theta_k \sim \pi_{\text{a}}(\cdot \mid \widehat{\boldsymbol{x}}_k)$, obtain $\widehat{\boldsymbol{x}}_{k+1} \sim \pi_t(\cdot \mid \widehat{\boldsymbol{x}}_k, \theta_k)$, and accrue r_k.

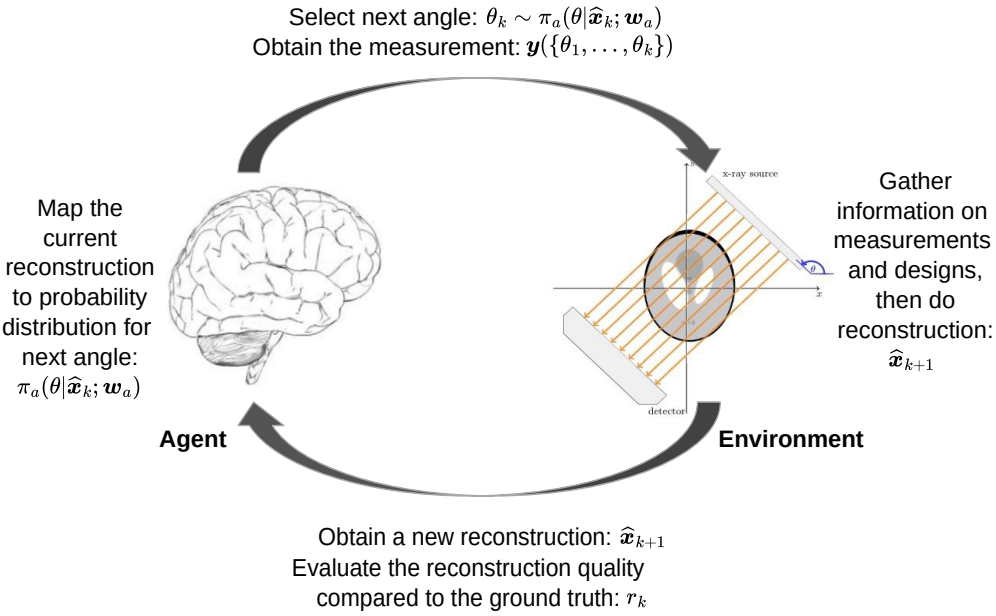


Figure 4.2: Acquisition–reconstruction loop at step k. **(1)** The current reconstruction (belief state) $\widehat{\boldsymbol{x}}_k$ is fed to the policy, which either terminates or selects the next angle θ_k. **(2)** If continuing, acquire the projection $\boldsymbol{y}(\theta_k)$. **(3)** Update the reconstruction with the new data to obtain the *updated* estimate $\widehat{\boldsymbol{x}}_{k+1}$. **(4)** Compute the reward r_k from the improvement in utility (PSNR to the ground truth $\bar{\boldsymbol{x}}$). Time flows left-to-right with the transition $k \to k+1$; no feedback "back to k" occurs. Notation: $\bar{\boldsymbol{x}}$ denotes the ground truth used solely for computing the training utility.

The learning objective is to maximize the expected return. For a fixed initial state $\widehat{\boldsymbol{x}}_1$,

$$J(\boldsymbol{w}_a) = V^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}_1). \tag{4.4}$$

For policy-gradient estimation we use the action–value and advantage functions. The action–value links actions to value via

$$V^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}) = \sum_{\theta} \pi_\mathrm{a}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)\, Q^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}, \theta), \tag{4.5}$$

where the action–value function is formally defined as

$$Q^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}_1, \theta) = \mathbb{E}_{\boldsymbol{\tau}^\theta \sim \pi^\theta_{\mathrm{chain}}} \left[\sum_{k=1}^{M} r_k \mid \widehat{\boldsymbol{x}}_1,\, \theta_1 = \theta \right], \tag{4.6}$$

and $\pi^\theta_{\mathrm{chain}}$ denotes the trajectory distribution conditioned on taking action θ at the first time step.

The advantage measures how much better an action is than the on-policy average:

$$A^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}, \theta) = Q^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}, \theta) - V^{\pi_\mathrm{a}}(\widehat{\boldsymbol{x}}), \tag{4.7}$$

which we use as a variance-reducing baseline in the policy gradient.

Building on this framework, we consider two approaches to optimal stopping. The key difference is whether stopping is modeled explicitly via a terminal policy $\pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$. Introducing π_{ter} augments the trajectory distribution in Equation (5.5) with an additional (Bernoulli) decision at each step, which in turn alters the policy–gradient contributions for the angle policy. In the terminal–policy variant, we also compute gradients with respect to $\boldsymbol{w}_t$.

4.3 Method

This section presents two approaches to optimal stopping. The first approach treats termination as an additional action in the action space. The second approach defines a separate termination policy, which is optimized jointly with the angle selection policy.

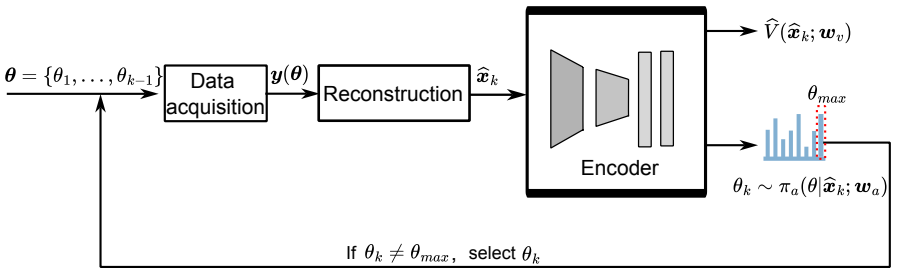
4.3.1 Naive optimal stopping

Optimal stopping for sOED can be implemented by introducing an additional terminal action within the action space. In our previous formulation [97], the action space consisted of the 180 discrete angles $0°, \ldots, 179°$. In the present work, we augment this space with an explicit termination action, yielding 181 discrete actions in total. The policy $\pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_{\mathrm{a}})$ produces logits $z(\widehat{\boldsymbol{x}}) \in \mathbb{R}^{181}$, inducing a categorical distribution $h = \mathrm{softmax}(z)$ over these actions. The selected action is the index $\theta \in \{1, \ldots, 181\}$, where $\theta = 1, \ldots, 180$ correspond to the angles $0°, \ldots, 179°$, and $\theta = 181$ denotes termination. The *reward function* for 'termination' and 'continuation' is defined by $R(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}}, \theta)$, which accounts for the decision to either continue or terminate, as follows:

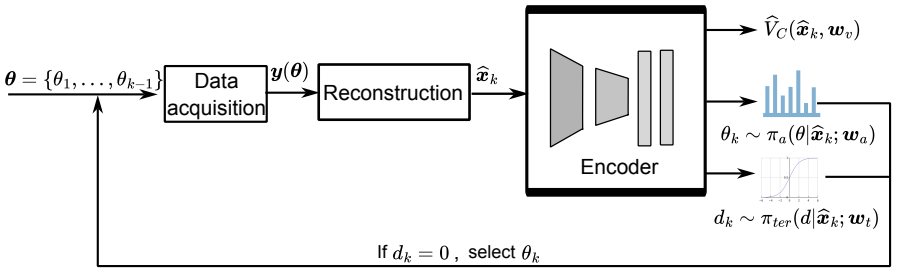
$$R(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}}, \theta) = \begin{cases} -b, & \text{if } \theta = \theta_{max}, \\ \mathrm{PSNR}(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}}), & \text{if otherwise,} \end{cases}$$

where $\mathrm{PSNR}(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}})$ is a function that serves as the immediate reward, evaluating the quality of the experiment at the stopping point $\widehat{\boldsymbol{x}}$, and $-b$ is a negative scaling factor representing the experimental cost incurred at each step. With PSNR (dB) as the utility, the per-step cost b is also in dB and serves as a minimum marginal-gain threshold (e.g., $b = 0.5$ means we continue only if PSNR increases by at least 0.5 dB); in applications b can be calibrated from time or dose per view. This mechanism works because the agent aims to maximize the cumulative reward: as long as selecting an additional angle leads to a PSNR gain that exceeds the experimental cost for continuing, the agent will proceed. It will choose to stop only when further actions are expected to result in no meaningful improvement in PSNR relative to the experimental cost. When termination is disabled (or $b = 0$), this approach reduces to the angle-only setting of [97].

As described in our previous work [97] and shown in Figure (4.3, a), after selecting $k-1$ angles $\boldsymbol{\theta} = \{\theta_1, ..., \theta_{k-1}\}$, the *observations* $\boldsymbol{y}(\boldsymbol{\theta})$ are obtained from the data acquisition. At step k, the reconstructed image $\widehat{\boldsymbol{x}}_k$ serves as the *belief state* and is fed to a shared encoder $E(\cdot)$; the initial state is $\widehat{\boldsymbol{x}}_1 = \mathbf{0}$. The encoder produces features $h_k = E(\widehat{\boldsymbol{x}}_k)$ that are consumed by Actor–Critic heads; no decoder is used, as images are reconstructed by the iterative solver. One branch is designed to estimate the state-value function $V^{\pi_{\mathrm{a}}}(\widehat{\boldsymbol{x}})$, and we use $\widehat{V}(\widehat{\boldsymbol{x}}; \boldsymbol{w}_v)$ for this approximation, where $\boldsymbol{w}_v$ denotes the neural network parameters. The other branch outputs the distribution over the *action space*, which consists of all possible angles and the terminal probability $\theta_{\max}$ as the final action. The inclusion of a mixture of actions for angles and termination



(a) Naive optimal stopping



(b) Optimal stopping using terminal policy

Figure 4.3: This figure illustrates the workflows: (a) Naive stopping: after selecting angles $\{\theta_1, \ldots, \theta_{k-1}\}$ and acquiring projections $\boldsymbol{y}(\boldsymbol{\theta})$, the reconstruction $\widehat{\boldsymbol{x}}_k$ becomes the belief state. A shared encoder then outputs (i) the state-value estimate $\widehat{V}(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$ and (ii) a distribution over all angles plus a terminal action $\theta_{\max}$. (b) Terminal-policy stopping: the encoder additionally outputs a probability distribution over termination versus continuation. Another distinction is that the top branch estimates the continuation value function using $\widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$. In both workflows, a termination signal halts the process. A convolutional encoder produces features consumed by Actor–Critic heads; no decoder is used, as images are reconstructed by the iterative solver.

represents the primary difference from the standard method described in our previous work [97]. Consequently, the Temporal Difference (TD) error [84] for the terminal action is modified to account for the reward at the termination state $\mathrm{PSNR}(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}})$, as shown in line 10 of Algorithm (2). Algorithm (2) outlines the modified algorithm based on the previous work [97].

Algorithm 2

1: Initialize the action policy parameters $\boldsymbol{w}_a$, and the value function parameters $\boldsymbol{w}_v$ randomly. Define the maximal experimental steps M. Set step sizes $\alpha^{\boldsymbol{w}_a} > 0$ and $\alpha^{\boldsymbol{w}_v} > 0$.
2: **for each episode do:**
3: Get a phantom sample $\bar{\boldsymbol{x}}$ then a zero matrix serves as the initial state $\widehat{\boldsymbol{x}}_1$ and $k = 0$.
4: **while:**
5: Select the angle based on the softmax policy, which maps the inputs
6: to a probability distribution that sums to 1: $\theta_k \sim \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$
7:
8: Get new measurements $\mathbf{y}_k$ from the data acquisition
9:
10: Reconstruct new image $\widehat{\boldsymbol{x}}_{k+1}$
11:
12: Get reward for continuation $-b$
13:
14: Estimate the state-values $\widehat{V}(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$ and $\widehat{V}(\widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_v)$
15:
16: Compute TD error:
17: **if** $\theta_k = \theta_{\max}$:
18: Get reward for termination $\mathrm{PSNR}(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}})$
19: $\delta_k = -b + \mathrm{PSNR}(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) - \widehat{V}(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$
20: **else:**
21: $\delta_k = -b + \widehat{V}(\widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_v) - \widehat{V}(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$
22:
23: Update the action policy function parameters $\boldsymbol{w}_a$:
24: $\boldsymbol{w}_a \leftarrow \boldsymbol{w}_a + \alpha^{\boldsymbol{w}_a} \nabla_{\boldsymbol{w}_a} \log \pi_{\mathrm{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \delta_k$
25: Update value function parameters $\boldsymbol{w}_v$:
26: $\boldsymbol{w}_v \leftarrow \boldsymbol{w}_v + \alpha^{\boldsymbol{w}_v} \nabla_{\boldsymbol{w}_v} \widehat{V}(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v) \delta_k$
27: Increase the step number k += 1
28:
29: **if** $\theta_k = \theta_{\max}$ **or** $k = M$:
30: **break**
31: **end for**

4.3.2 Optimal stopping using terminal policy

Objective function

Different from the naive way of implementing optimal stopping, we consider an independent parameterized terminal policy [5], which maps the state $\widehat{\boldsymbol{x}}$ to a stochastic stopping decision $d \sim \pi_{\text{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$. The continuation corresponds to $d = 0$, while termination corresponds to $d = 1$. Apart from the stopping decision, the reward function remains the same as in naive optimal stopping.

In the context of optimal stopping for sOED, the framework integrates both the terminal policy for scan termination and the action policy for angle selection, instead of combining them within the same action space as in the naive approach. First, we reformulate the probability chain from Equation (5.5) to incorporate the termination decision. The updated probability chain reflects the decision-making process of the terminal policy, which evaluates at each step whether the trajectory should terminate or continue.

Similar to Equation (5.5), the probability chain $\pi_{\text{chain,C}}(\boldsymbol{\tau}_C; \boldsymbol{w}_a, \boldsymbol{w}_t)$ with the inclusion of the termination distribution is expressed as:

$$\pi_{\text{chain,C}}(\boldsymbol{\tau}_C; \boldsymbol{w}_a, \boldsymbol{w}_t) = \prod_{k=1}^{T} \pi_{\text{sec}}(d_k, \widehat{\boldsymbol{x}}_{k+1}, \theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a, \boldsymbol{w}_t), \tag{4.8}$$

where the function π_{sec} is defined as:

$$\pi_{\text{sec}}(d, \widehat{\boldsymbol{x}}', \theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a, \boldsymbol{w}_t) = \pi_{\text{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t) + \Big(1 - \pi_{\text{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)\Big) \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a) \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta). \tag{4.9}$$

The stopping step T is determined by the terminal policy, which is less or equal to the maximal step M and $d_M \equiv 1$.

Convention. The terminal policy is Bernoulli over $d \in \{0, 1\}$, with $d = 1$ denoting termination and $d = 0$ continuation. Whenever d appears inside $\pi_{\text{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$ without an explicit sum over d, it refers to the event $d = 1$. Consequently, $1 - \pi_{\text{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$ denotes the probability of continuation.

A trajectory up to the stopping point is generated from the new $\pi_{\text{chain,C}}$ as:

$$\{\widehat{\boldsymbol{x}}_1, (d_1, \theta_1, \widehat{\boldsymbol{x}}_2, -b), (d_2, \theta_2, \widehat{\boldsymbol{x}}_3, -b), \ldots, (d_T, \theta_T, \widehat{\boldsymbol{x}}_{T+1}, p_T)\}.$$

This trajectory represents the sequence of states ($\widehat{\boldsymbol{x}}$), termination indicators (d),

experimental cost ($-b$), and quality evaluations $\mathrm{PSNR}(\widehat{\boldsymbol{x}}_T, \bar{\boldsymbol{x}})$ (p_T) at termination. The process is designed to terminate at step T as soon as $d_T = 1$, ensuring that the trajectory is finite and explicitly concludes. For all prior steps ($i < T$), $d_i = 0$, indicating that the process continues during those steps.

The objective function, representing the value function at the fixed initial state $\widehat{\boldsymbol{x}}_1$, aims to learn both the action and terminal policies that maximize the expected cumulative rewards:

$$\begin{aligned} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \\ &= \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\, \mathrm{PSNR}(\widehat{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}) \\ &\quad + \Big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\Big)\, V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1). \end{aligned} \tag{4.10}$$

Here, the continuation state-value function is defined as follows [23, 63]:

$$V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) = -b + \mathbb{E}_{\boldsymbol{\tau}_C^{(2)} \sim \pi_{\mathrm{chain,C}}^{(2)}} \left[V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_2) \right], \tag{4.11}$$

where $\boldsymbol{\tau}_C^{(2)}$ represents the trajectory starting from $\widehat{\boldsymbol{x}}_2$, following the probability chain described in Equation (4.8).

Similarly, the continuation action-value function is defined only when the trajectory continues, as one of its inputs is the action: $Q_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta)$.

The optimal terminal policy $\pi_{\mathrm{ter}}^*(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)$ at step k is defined as [23, 63, 11]:

$$\pi_{\mathrm{ter}}^*(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t) = \begin{cases} \mathbb{I}\Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}}) \geq V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_k)\Big), & \text{if } k < M, \\ 1, & \text{if } k = M, \end{cases} \tag{4.12}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and M is the maximum number of steps in the experiment.

Figure (4.3, b) illustrates the workflow of optimal stopping using the terminal policy. Compared to the naive optimal stopping, this approach includes an additional branch for the terminal policy, which uses the sample from Sigmoid function to determine continuation or termination. Another distinction is that the top branch estimates the continuation value function $V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_k)$ using $\widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$.

Policy gradient

To jointly solve the optimal stopping and action selection problems for sOED, we propose a novel policy gradient method. The policy gradients are calculated using

Equation (4.10).

For the gradient with respect to the action policy $\boldsymbol{w}_a$, the detailed unrolling recursive derivation is provided in Appendix (B.2). By sampling from N trajectories, the gradient is expressed as:

$$\begin{aligned}\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= \nabla_{\boldsymbol{w}_a} V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \\ &\propto \sum_{n=1}^{N} \left(\sum_{k=1}^{T} \nabla_{\boldsymbol{w}_a} \log \pi_{\mathrm{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) Q_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_k, \theta) \right),\end{aligned} \tag{4.13}$$

To improve the stability and efficiency of action policy gradient computation, a baseline is introduced by replacing the continuation action-state value function in Equation (4.13) with an advantage function [84]. The continuation advantage function is defined by incorporating the output of the terminal policy; further details are provided in the Appendix (B.1). The continuation advantage function is approximated as:

$$\begin{aligned}A_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta) \approx -b &+ \Big(1 - \pi_{\mathrm{ter}}(d^{'} \mid \widehat{\boldsymbol{x}}^{'}; \boldsymbol{w}_t)\Big) V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}^{'}) \\ &+ \pi_{\mathrm{ter}}(d^{'} \mid \widehat{\boldsymbol{x}}^{'}; \boldsymbol{w}_t) \mathrm{PSNR}(\widehat{\boldsymbol{x}}^{'}, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}).\end{aligned} \tag{4.14}$$

We perform stochastic gradient ascent to update the parameters of the angle policy, $\boldsymbol{w}_a$, using a step size of $\alpha^{\boldsymbol{w}_a}$.

$$\boldsymbol{w}_a \leftarrow \boldsymbol{w}_a + \alpha^{\boldsymbol{w}_a} \nabla_{\boldsymbol{w}_a} \log \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a) A_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta) \tag{4.15}$$

For the gradient with respect to the terminal policy $\boldsymbol{w}_t$, the detailed unrolling recursive derivation is provided in the Appendix (B.3). By sampling from N trajectories, the gradient is expressed as:

$$\begin{aligned}\nabla_{\boldsymbol{w}_t} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= \nabla_{\boldsymbol{w}_t} V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) \\ &\propto \sum_{n=1}^{N} \left(\sum_{k=1}^{T} \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t) \Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_k) \Big) \right),\end{aligned} \tag{4.16}$$

We also perform stochastic gradient ascent to update the parameters of the terminal policy, $\boldsymbol{w}_t$, using a step size of $\alpha^{\boldsymbol{w}_t}$.

$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_t + \alpha^{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t) \Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) \Big) \tag{4.17}$$

The complete algorithm is presented in Algorithm (3).

Algorithm 3

Initialize the action policy parameters $\boldsymbol{w}_a$, the terminal policy parameters $\boldsymbol{w}_t$, and the value function parameters $\boldsymbol{w}_v$ randomly. Define the maximal experimental steps M. Set step sizes $\alpha^{\boldsymbol{w}_a} > 0$, $\alpha^{\boldsymbol{w}_t} > 0$, $\alpha^{\boldsymbol{w}_v} > 0$.
for each episode do:
    Get a phantom sample $\bar{\boldsymbol{x}}$ then a zero matrix serves as the initial state $\widehat{\boldsymbol{x}}_1$ and $k = 0$.
    while $d_k = 0$ **and** $k < M$**:**
        Select the angle based on the softmax policy, which maps the inputs to a probability distribution that sums to 1: $\theta_k \sim \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$
        Get new measurements $\mathbf{y}_k$ from the data acquisition

        Reconstruct new image $\widehat{\boldsymbol{x}}_{k+1}$

        Get reward for continuation $-b$, and rewards for termination PSNR$(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}})$ and PSNR$(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}})$

        Determine the terminal action based on the Sigmoid policy, which maps the inputs to terminal probability: $d_k \sim \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)$
        Estimate the state-values $\widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$ and $\widehat{V}_C(\widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_v)$

        Compute TD error: $\delta_k = -b + \Big(1 - \pi_{\mathrm{ter}}(d_{k+1} \mid \widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_t)\Big)\widehat{V}_C(\widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_v)$
        $+ \pi_{\mathrm{ter}}(d_{k+1} \mid \widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_t)\mathrm{PSNR}(\widehat{\boldsymbol{x}}_{k+1}, \bar{\boldsymbol{x}}) - \widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$

        Update policy function parameters $\boldsymbol{w}_a$:
        $\boldsymbol{w}_a \leftarrow \boldsymbol{w}_a + \alpha^{\boldsymbol{w}_a} \nabla_{\boldsymbol{w}_a} \log \pi_{\mathrm{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\delta_k$
        Update policy function parameters $\boldsymbol{w}_t$:
        $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_t + \alpha^{\boldsymbol{w}_t} \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)\Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}}) - \widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)\Big)$
        Update value function parameters $\boldsymbol{w}_v$:
        $\boldsymbol{w}_v \leftarrow \boldsymbol{w}_v + \alpha^{\boldsymbol{w}_v} \nabla_{\boldsymbol{w}_v} \widehat{V}_C(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)\delta_k$
        Increase the step number k += 1

end for

Termination is modeled either as (Algorithm 2) an additional action in a single 181-way categorical policy, or (Algorithm 3) a dedicated binary termination head alongside a 180-way angle head, both sharing the same encoder. In both cases, policy parameters are optimized jointly without alternating/block-coordinate updates. Unlike exhaustive search, which evaluates many remaining candidates at each step, the learned policy selects one angle per-step. Once trained offline, the model can be used directly online;

per-step latency is effectively dominated by the reconstruction update.

4.4 Results

The proposed optimal stopping method for sOED was validated through a series of X-ray CT experiments. The model was first trained using synthetic data and then evaluated on experimental X-ray CT data to test its performance and generalizability.

4.4.1 Dataset

Synthetic dataset

We created a synthetic dataset that includes features with strongly nonuniform orientations, corresponding to highly informative imaging angles. The dataset includes three shapes—parallelograms, triangles, and pentagons—each varying in scale, rotation, and position. These variations result in differing requirements for the number of angles needed to achieve accurate reconstructions. Figure (4.4) provides example images from this dataset, while the parameters used for its generation are outlined in the Appendix (B.5).

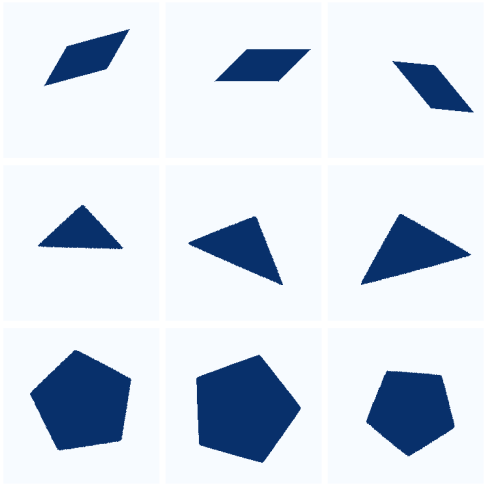


Figure 4.4: The figure shows nine samples from the synthetic dataset, including parallelograms, triangles, and pentagons. Each shape type is represented by three samples.

Experimental dataset

All data were collected at the FleX-ray laboratory of Centrum Wiskunde en Informatica (CWI) in Amsterdam, the Netherlands [98, 48]. A scanning approach was used to acquire projections with dimensions of 956×10 pixels. The source-to-object and detector-to-object distances were both set to 225 mm. An exposure time of 80 ms per-projection was applied, and the source spectrum was shaped using filters consisting of 0.1 mm zinc, 0.2 mm copper, and 0.5 mm aluminum. A total of 3601 projections were acquired. Figure (4.5) illustrates the scanning setup.

The dataset comprises two laser-cut objects—triangular and pentagonal—fabricated from 6 mm-thick transparent acrylate. Each shape is represented by 12 samples of varying sizes. The right-angle edges of the triangular samples range from 2.8 cm to 4.0 cm, while the edges of the pentagonal samples range from 2.5 cm to 3.0 cm. Each scanning session involved different placements of the objects, resulting in variations in rotation and translation. To create a dataset with two noise levels, two emission currents were used: 600 μA and 100 μA. The lower emission current (100 μA) produced data with higher noise levels. Finally, 12 groups of projections were acquired for each shape and each noise level, resulting in a total of 48 groups of projections.

For the purpose of this study, the number of projections was reduced to 361 by subsampling every 10th projection. To approximate the cone-beam geometry as a fan-beam geometry, only the middle row of the detector was used. The column size of the projections was then reduced to 239 by selecting every fourth pixel on the detector. Furthermore, the fan-beam data was rebinned into a parallel-beam data with 180 projections to simplify angle selection. This transformation made it more precise to identify the informative angles, as they are tangential to the edges in parallel-beam geometry.

4.4.2 Implementation

For training on the synthetic data, the Astra Toolbox [90, 89] was used to generate the simulated projections. Reconstruction $\widehat{\boldsymbol{x}}(\boldsymbol{\theta})$ is computed with SIRT under a nonnegativity constraint (150 iterations), a stable choice for sparse-view CT. The reconstructor is held fixed throughout to isolate the effect of the design policy—angle selection and stopping—which is the focus of this study.

For the algorithm implementation, the architectures of the encoder and the Actor–Critic neural networks are detailed in the Appendix (B.6). During training, weights of 1.0 and 0.5 were assigned to the actor loss and critic loss in Algorithm (2) and

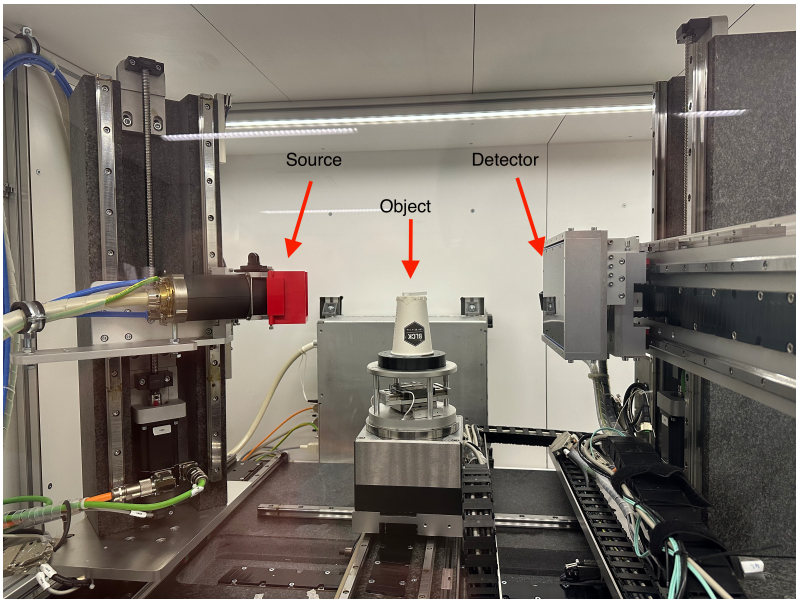


Figure 4.5: Experimental scanning setup at the FleX-ray laboratory. The X-ray source (left) and flat-panel detector (right) remain fixed, while the object is positioned on a paper cup atop the rotation stage centered between them. During acquisition, the stage rotates to capture projections from multiple angles.

Algorithm (3,) respectively. To encourage exploration during training, an entropy loss with a weight of 0.01 was included. Compared to Algorithm (2), Algorithm (3) included an additional loss term for the terminal policy with a weight of 1.0. These parameter settings were empirically chosen to strike an optimal balance between policy optimization, accurate value estimation, and robust exploration during training. The network weights were optimized using the Adam optimizer [47] with a learning rate of 10^{-4} and a weight decay of 10^{-5}. To prevent selecting the same angle multiple times, previously chosen angles are masked according to the procedure described in [44].

4.4.3 Training on synthetic data with Gaussian noise

In this experiment, we investigated whether the two algorithms could adaptively determine the optimal number of angles for CT imaging by adjusting the experimental costs before termination. The projections incorporated 5% Gaussian noise. For Algorithms (2) and (3), the experimental costs ranged from -0.4 to -0.9 in intervals of 0.1. At higher experimental costs, priority is placed on guaranteeing image quality, whereas at lower costs the emphasis shifts toward reducing the number of projections. A separate model was trained for each experimental cost setting using the synthetic dataset shown in Figure (4.4). To prevent excessive nontermination during training, the maximum number of angles, M, in Algorithm (2) and Algorithm (3), was limited to 20.

During training, each episode involved sampling a data point from the dataset shown in Figure (4.4). The policy was trained at each angle selection step until termination, thereby completing the episode. A total of 80,000 episodes (i.e., 80,000 sampled data) were considered. Figure (4.6), and Figure (4.7)) report training dynamics on the training distribution (5% Gaussian noise) and are used to analyze learning behavior. Quantitative evaluation on held-out data appears in Section 5.4 (unseen rotations and noise levels) and Section 5.5 (experimental X-ray CT).

However, the naive stopping mechanism in Algorithm (2) (stop-as-action) did not function reliably. Across per-step cost settings $b \in [0.4, 0.9]$ (displayed in the plots as $-b$), the policy typically exhausted the maximum number of angles. As shown in Figure (4.6), the setting $b = 0.5$ (label "-0.5") produced shape-dependent stopping with different angle counts, whereas $b = 0.6$ (label "-0.6") failed to yield a valid stopping policy. These outcomes indicate that, in this formulation, the policy-gradient signal to the "stop'' class is too weak/sparse, leading to instability, lack of robust convergence, and a tendency to ignore the terminal action.

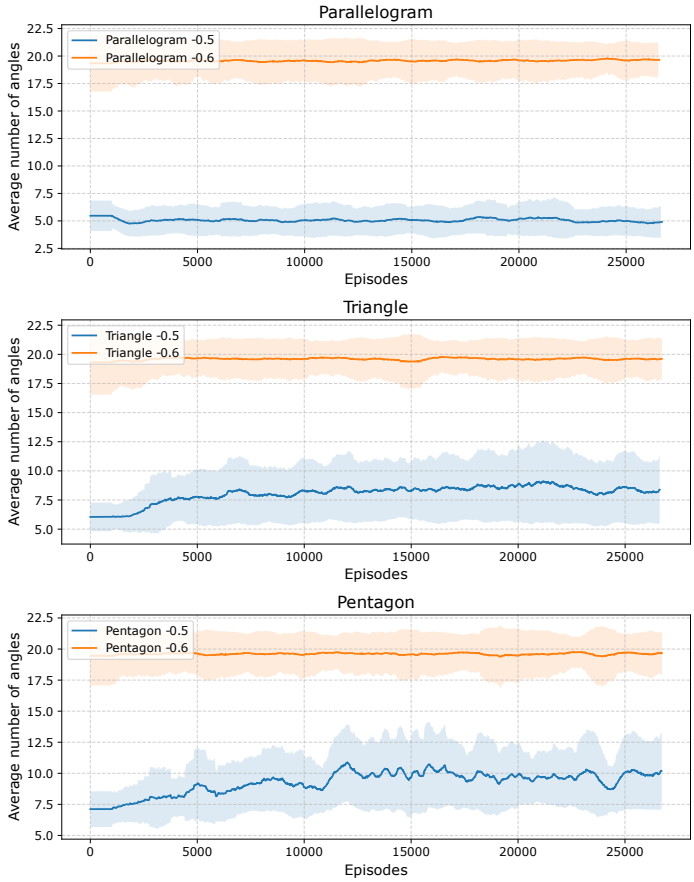


Figure 4.6: Comparison of the number of angles selected by the naive policy and by experimental cost settings of -0.5 and -0.6. The curves represent the mean number of angles during training, averaged over every 1000 episodes, and the shaded regions indicate the variance. Results are grouped by shape: parallelogram, triangle, and pentagon.

In contrast, the terminal policy remained robust under different reward settings. Figure (4.7) shows that, during training, the number of angles increased, and more complex shapes required more angles. Lower costs (experimental cost is -0.5) generally led to a higher number of angles.

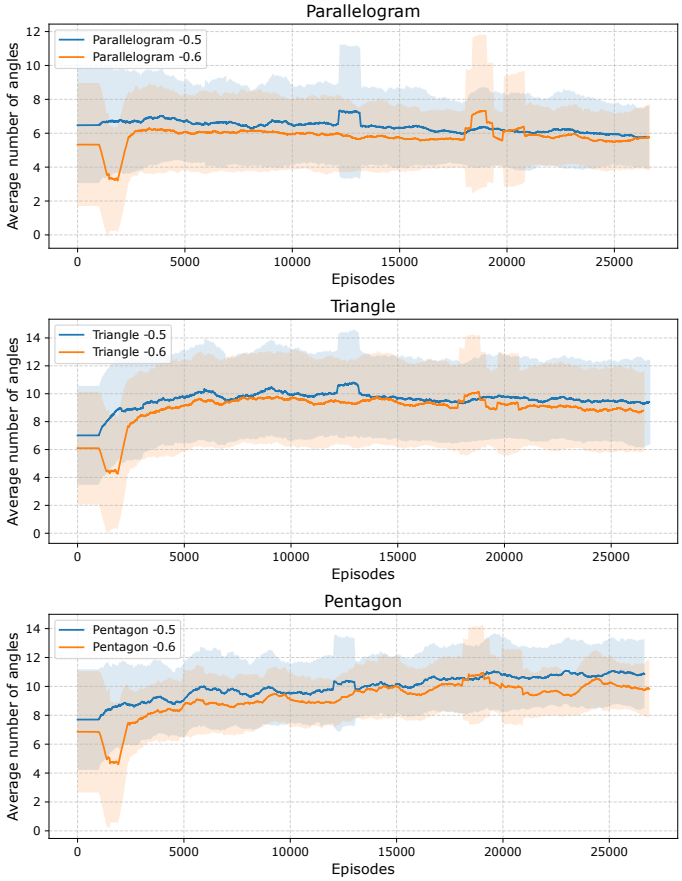


Figure 4.7: Comparison of the number of angles selected by the terminal policy and by experimental cost settings of -0.5 and -0.6. The curves represent the mean number of angles during training, averaged over every 1000 episodes, and the shaded regions indicate the variance. Results are grouped by shape: parallelogram, triangle, and pentagon.

4.4.4 Validation on unseen rotations and noise levels

To assess the generalizability of Algorithm (3) for optimal stopping, we applied the policy trained in the previous experiment (which involved 5% Gaussian noise) to syn-

thetic data featuring unseen rotations—rotations not included in the training set shown in Figure (4.4)—as well as two additional noise levels: 3% and 7% Gaussian noise. The total number of unseen phantoms was 1,800, with 600 for each shape.

A standard baseline approach, the *Golden Ratio (GR) Policy* [49, 21], is considered for comparison with the RL policy. In the GR policy, an angular increment based on an irrational number is used, which leads to a nonrepeating sequence of angles that fill the angular space most evenly over time. We note that simple online stopping heuristics—e.g., stopping when the relative data residual or the relative reconstruction change falls below a threshold for L consecutive steps—require dataset-specific tuning and can either stop too early or over-acquire. Instead, we adopt a learned terminal policy that internalizes the quality–cost trade-off via the per-step cost b.

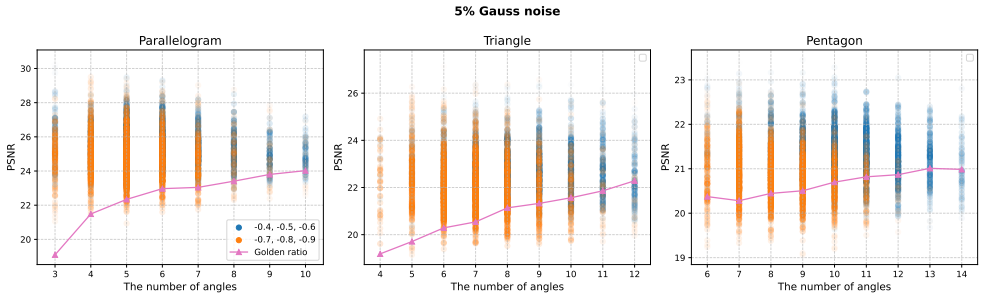


Figure 4.8: Simulation results are categorized into parallelogram, triangle, and pentagon shapes to illustrate how the number of angles influences PSNR under various rewards and various noise levels. Data points are color-coded by experimental costs—blue $(-0.4, -0.5, -0.6)$ and orange $(-0.7, -0.8, -0.9)$. Increased transparency indicates data points that are further from the mean. Triangular markers represent the mean PSNR values obtained from the golden ratio policy at each corresponding angle.

Figure (4.8) shows the validation results for experimental costs ranging from -0.4 to -0.9 (in increments of 0.1) under Gaussian noise level of 5%. Results for Gaussian noise levels of 3% and 7% are presented in Appendix (B.4). Each data point represents the relationship between the number of angles and the corresponding PSNR values. For clarity, results are grouped by shape (parallelogram, triangle, and pentagon). The GR policy used the same synthetic data with unseen rotations and the same number of angles chosen by the terminal policy. For each number of angles used by the GR policy, the mean PSNR value was calculated.

Overall, the RL approach with the terminal policy achieved higher PSNR values compared to the GR policy. At lower numbers of angles, the GR policy consistently produced lower mean PSNR values across all shapes, with the difference between the

Table 4.1: Comparison of the number of angles and PSNR values for triangle and pentagon shapes under different noise levels.

Shape	Noise	Number of angles	RL (PSNR)	GR (PSNR)
Parallelogram	3%	6.02 ± 2.26	25.96 ± 1.29	22.65 ± 2.17
Parallelogram	5%	6.96 ± 2.81	25.55 ± 1.10	22.92 ± 2.06
Parallelogram	7%	9.41 ± 3.98	24.99 ± 0.95	23.21 ± 1.91
Triangle	3%	8.54 ± 1.97	23.42 ± 1.10	21.58 ± 1.71
Triangle	5%	9.43 ± 2.21	22.85 ± 1.09	21.44 ± 1.54
Triangle	7%	11.39 ± 3.09	22.05 ± 1.12	21.17 ± 1.41
Pentagon	3%	10.43 ± 1.98	22.66 ± 0.67	22.21 ± 0.51
Pentagon	5%	10.99 ± 2.01	21.47 ± 0.56	20.99 ± 0.51
Pentagon	7%	12.56 ± 2.25	20.01 ± 0.58	19.56 ± 0.52

two methods diminishing as the number of angles increased. As shape complexity increased from parallelogram to pentagon, the performance gap between the two policies narrowed. For instance, at lower numbers of angles for the parallelogram, the RL approach yielded data points above the mean value of the GR policy, while for the more complex pentagon shape, the RL policy's advantage was less pronounced. Figure (4.8) uses distinct colors to represent different ranges of experimental costs, with generally lower experimental costs leading to fewer angles. As experimental cost increases, the number of angles decreases across all shapes. In Figure (4.8), the orange points (indicating higher costs) are concentrated at lower numbers of angles, while the blue points (indicating lower costs) are found at higher numbers of angles; however, the distinction is more evident for complex shapes, which require more angles.

Table (4.1) summarizes the number of angles selected for the three shapes during validation under different noise levels, with the experimental cost fixed at -0.5. In general, lower noise levels lead to fewer angles being chosen. As shown in the table, each shape requires the fewest angles at 3% Gaussian noise, whereas 7% Gaussian noise leads to the largest number of angles. Furthermore, the number of angles increases with the complexity of the geometry: the parallelogram requires the fewest angles, while the pentagon requires the most. The RL policy outperforms the GR policy under every tested condition.

4.4.5 Test on experimental X-ray CT data with two noise levels

We explored the gap between simulation and the real world by applying the model trained on the synthetic dataset to the experimental X-ray CT data described in Section 5.1.2. Building on the potential demonstrated in the simulation, we further investigated whether the trained model could adapt to changes in the real scanning environment, particularly variations in noise levels. Two noise levels were considered, based on different emission current settings (600 μA and 100 μA). The model was trained using 5% Gaussian noise on the projections, while the noise in the experimental X-ray CT typically consists of a mixture of Poisson and Gaussian noise [2]. It is important to demonstrate whether training under simplified conditions (Gaussian noise model) can produce reasonable results when applied to the more complex conditions of the real world (experimental noise).

Figure (4.9) shows results obtained from experimental X-ray CT data using three trained models under varying emission currents and reward settings. To evaluate reconstruction quality, we used the reconstruction obtained with all 180 angles as the ground truth. The GR policy used the same number of angles chosen by the terminal policy. For each number of angles used by the GR policy, the mean PSNR value was calculated. Several consistent trends emerged across the 12 data groups, aligning with the conclusions from the validation. First, pentagon data points generally involved a larger number of selected angles than triangles. Second, noisier data from the lower emission current (100 μA) triggered more angles than data collected at the higher emission current (600 μA). Third, although the number of samples was limited, it was still evident that lower experimental costs (e.g., -0.4, -0.5 and -0.6) led to more angles, while higher costs (e.g., -0.7, -0.8 and -0.9) required fewer angles. Finally, the RL policy clearly outperformed the GR policy when fewer angles were selected, though the difference between them diminished as the number of angles increased.

Table (4.2) compares the performance of the RL policy and GR policy under an emission current of 600 μA and 100 μA under a reward setting of -0.5. The RL policy consistently outperformed the GR policy for triangles, while it was only comparable to the baselines for pentagons. This outcome, likely influenced by the gap between synthetic and experimental X-ray CT data, is consistent with the training results, where the RL policy did not demonstrate a particularly distinct advantage for pentagon shapes.

Additionally, Figure (4.10) illustrates sample reconstructions for the two policies, with the number of angles determined by the RL policy. Each colored spoke indicates

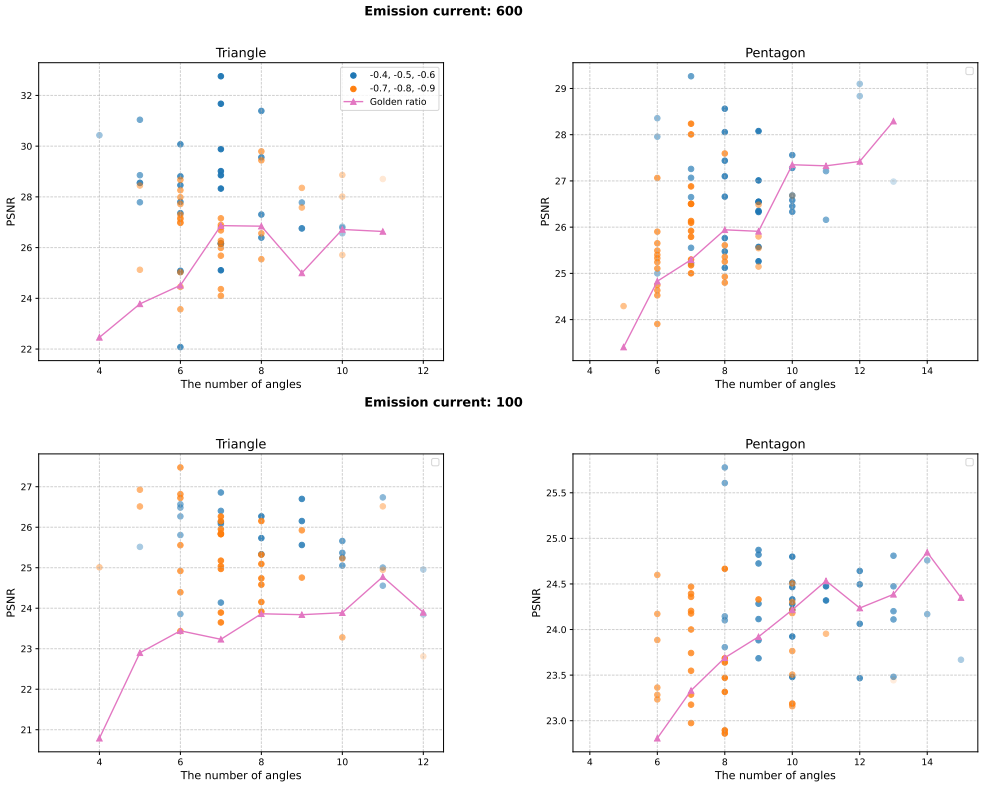


Figure 4.9: Results on experimental X-ray CT data for the triangle and pentagon shapes are shown, illustrating how the number of angles affects PSNR under various experimental costs and noise levels. Data points are color-coded by their experimental costs: blue $(-0.4, -0.5, -0.6)$ and orange $(-0.7, -0.8, -0.9)$. Triangular markers represent the mean PSNR values obtained from the GR policy at each corresponding number of angles.

Table 4.2: Comparison of the number of selected angles and PSNR values for triangle and pentagon shapes under two different noise levels: The low noise level corresponds to an emission current of 600 μA, and the high noise level to 100 μA.

Shape	Noise	Number of angles	RL (PSNR)	GR (PSNR)
Triangle	Low	7.92 ± 3.04	27.38 ± 2.09	25.58 ± 2.10
Triangle	High	11.08 ± 2.87	25.95 ± 0.72	24.47 ± 1.32
Pentagon	Low	9.17 ± 1.86	26.63 ± 1.13	26.84 ± 0.87
Pentagon	High	10.33 ± 2.05	24.50 ± 0.63	24.40 ± 0.30

the *beam direction* of a parallel-beam view at angle θ (degrees), measured counterclockwise from the x-axis; note that θ and $\theta + 180^\circ$ are equivalent. The results show that the angles selected by the RL policy tend to cluster around the edges. Furthermore, as the noise level increases (from an emission current of 600 μA to 100 μA), the number of selected angles increases, and their distribution becomes broader.

Consequently, the RL policy trained on synthetic data with a simple Gaussian noise model demonstrated reasonable performance on experimental X-ray CT data, effectively handling both optimal stopping and the selection of informative angles.

4.5 Discussion

These results demonstrate that our algorithm effectively balances experimental costs and reconstruction quality by incorporating optimal stopping and adaptively selecting informative angles through deep reinforcement learning. Moreover, the model trained on a synthetic dataset with a simple Gaussian noise model showed satisfactory performance on experimental X-ray CT data, which typically includes more complex noise, such as a mixture of Gaussian and Poisson noise. The algorithm also demonstrated its ability to adapt the number of angles (optimal stopping) to varying noise levels in real-world scenarios while maintaining a consistent level of experimental quality.

Despite these encouraging findings, several areas for improvement remain. First, the gap between simulation and real-world performance could be minimized by incorporating more realistic simulators that account for precise noise models. Second, the training process could be improved by adopting a multi-stage approach, where the model is initially trained on simulated data and subsequently fine-tuned using experimental X-ray CT data. Third, extending the dataset to three-dimensional scenarios would enable the exploration of larger action and state spaces, further enhancing the model's capability. Additionally, task-specific zooming techniques could be integrated for applications such as defect detection. Fourth, incorporating learning-based reconstruction algorithms could facilitate the extraction of informative features directly from the reconstruction process, providing a more robust basis for selecting informative angles. Fifth, while we report PSNR as the primary, model-aligned metric in this work, complementary structural and task-oriented measures (e.g., SSIM, edge preservation, and detectability) are also informative; related results with SSIM and detectability are reported in [96], and a comprehensive multi-metric evaluation is a natural direction for future work. Finally, the learned policy is trained on the shape families used in this work and generalizes within that distribution (unseen rotations/scales); we do not

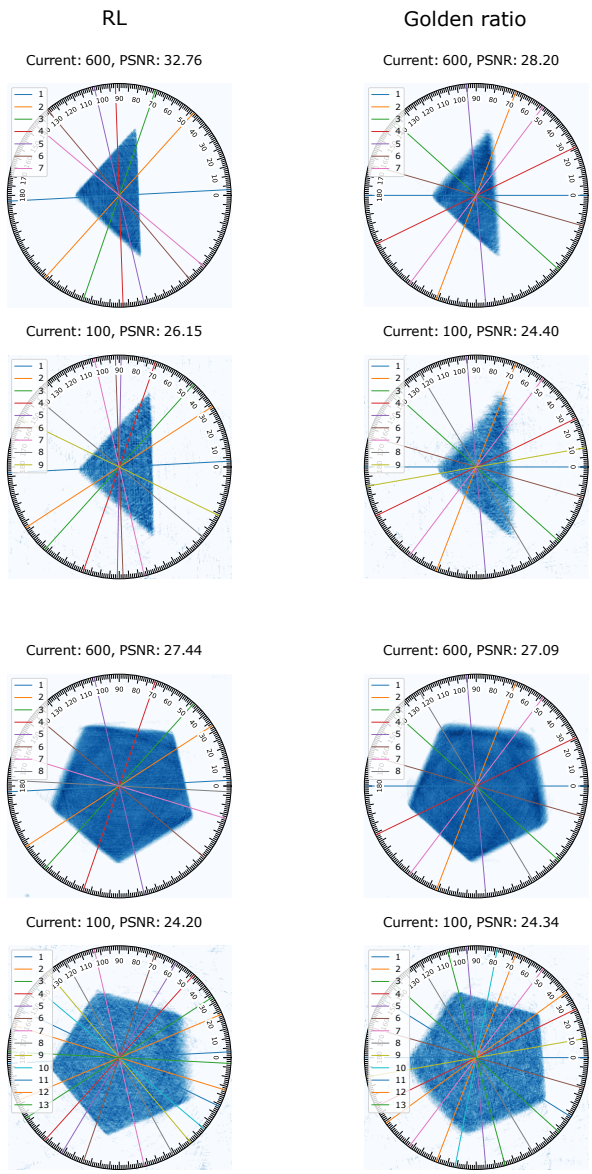


Figure 4.10: Selected projection directions for a representative scan. Each colored ray denotes the projection *normal* (central ray) of a parallel-beam view at angle θ (degrees), measured counter-clockwise from the x-axis and defined modulo 180°. Color encodes acquisition order (step $1 \to M$; see colorbar).

claim out-of-distribution generalization to arbitrary non-convex or non-homogeneous objects. In industrial CT, CAD/blueprint priors make class-specific training natural; extending to broader multi-material classes is left for future work.

4.6 Conclusion

In this paper, we proposed an approach to simultaneously optimize adaptive informative angle selection and optimal stopping by introducing a terminal policy and jointly computing the policy gradient for both the angle selection and terminal policies. Additionally, we investigated the gap between simulation and real-world scenarios. Our findings demonstrate the feasibility of achieving optimal stopping for sOED based on experimental costs. Furthermore, the trained model from simulation showed promising potential for application to experimental X-ray CT data, highlighting its value for industrial CT applications. This approach paves the way for fully adaptive scanning processes, optimizing both the selection of informative angles and the number of angles required.

4.7 Data and code availability

Experimental datasets are available on Zenodo: `https://zenodo.org/records/14893740`

The code for this work is available on GitHub: `https://github.com/tianyuan1wang/Optimal_Stopping_RL_CT`

Chapter 5

Deep Reinforcement Learning for Optimizing Angle Selection and Dose Allocation in CT Reconstruction

5.1 Introduction

X-ray Computed Tomography (CT) is a widely used non-destructive imaging modality that reconstructs the internal structure of objects from external projections. However, CT inherently faces trade-offs among image quality, scan time, and radiation exposure —particularly in settings requiring rapid inspection or involving dose-sensitive subjects.

In medical CT, *dose optimization* is critical to minimize patient radiation exposure. As reviewed in [55], conventional approaches such as tube current modulation and automatic exposure control adjust exposure according to patient size and

This chapter is based on:

Tianyuan Wang, Daniël M Pelt, Felix Lucka, Tristan van Leeuwen, and Kees Joost Batenburg. "Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction". In: *Applied Mathematics in Science and Engineering* 34.1 (2026), p. 2705107. DOI: 10.1080/27690911.2026.2705107

anatomy. Although effective, these strategies remain largely heuristic and lack task-specific adaptivity. In *industrial CT*, the objective is often to concentrate measurements on regions with high uncertainty or suspected defects while avoiding redundant projections. Similar considerations arise in computational microscopy and electron or optical tomography, where optimizing illumination dose is essential to prevent specimen damage. The object's position or orientation may differ between scans due to manipulation or changes in setup. These factors motivate the development of *adaptive scanning strategies* that jointly optimize projection angle selection and dose allocation under a fixed dose budget. Depending on the imaging task, the total dose can be distributed over many low-intensity views or concentrated on fewer, more informative angles, since not all measurements contribute equally to reconstruction quality [46].

Several works investigate adaptive scanning strategies that focus solely on sequential angle selection, where projection views are chosen based on information acquired during the scan [9, 26, 7, 97]. The problem of selecting informative projection angles has often been formulated as a *bi-level optimization* problem [66], where the lower level performs image reconstruction, and the upper level minimizes reconstruction error. This approach is high-dimensional and non-convex, as both levels depend on the design parameters (angles). Alternatively, *Bayesian experimental design* [53, 67, 64] seeks to maximize expected information gain by comparing prior and posterior distributions or reconstruction accuracy. Extensions of this framework have incorporated prior knowledge to guide angle selection and improve acquisition efficiency [8, 9, 22, 39, 7, 26]. However, extending these methods to include *dose allocation* further increases dimensionality and computational cost. In practical applications, imaging often requires adaptive, online decision-making to accommodate variations in object geometry or motion. Traditional Bayesian approaches, which rely on repeated evaluation of expected information gain, are computationally intensive and unsuitable for real-time control. Moreover, the *joint optimization of angle and dose* remains largely unexplored, despite its strong impact on reconstruction quality.

Recent works have explored *Reinforcement Learning (RL)* for sequential CT scanning [78, 97, 96, 99]. Policy-based RL methods learn scanning strategies directly from data, eliminating repeated optimization during acquisition and enabling adaptive decision-making based on intermediate reconstructions. Once trained, these policies can be applied efficiently in online scanning scenarios.

From a reconstruction perspective, the *incident photon count*—that is, the local dose—directly affects measurement noise according to the Beer–Lambert law. When dose allocation varies across angles, high-dose projections yield lower noise, while

low-dose projections are more uncertain. This heterogeneity calls for reconstruction algorithms that explicitly account for dose-dependent noise. *Penalized Weighted Least Squares (PWLS)* methods [95, 45] address noise in low-dose CT but typically assume uniform dose across angles. Some learning-based approaches incorporate global dose levels [101], yet they do not handle unequal per-angle dose distributions.

In this work, we address this gap by deriving a *dose-aware PWLS* model derived from a second-order Taylor approximation of the Poisson log-likelihood. While conventional PWLS formulations generally assume uniform or fixed doses across views, our adaptive setting requires handling intentionally non-uniform, per-angle dose allocations. To accommodate this, we introduce an angle-dependent weighting matrix that explicitly reflects the dose assigned to each view by the adaptive policy.

Contributions

Building on recent advances in adaptive imaging, this work presents an RL-based framework for joint projection-angle selection and dose allocation in X-ray CT. The main contributions are:

- **Joint angle–dose acquisition strategy:** We formulate adaptive CT acquisition as a sequential decision-making problem in which the policy selects projection angles under a fixed dose budget. Dose allocation is represented through repeated selections of projection angles, so that more photons are assigned to views selected multiple times. The resulting policy is trained with task-dependent objectives, including reconstruction quality and defect detectability.

- **Dose-aware reconstruction under non-uniform exposure:** We extend the PWLS reconstruction model to incorporate *angle-dependent weights* induced by the non-uniform dose distribution produced by the acquisition policy. The resulting dose-weighted PWLS data-fidelity term is combined with a Total Variation (TV) prior within a Plug-and-Play (PnP) reconstruction scheme, enabling reconstruction in settings where the statistical reliability of the measured projections varies across angles.

The remainder of this paper is organized as follows. Section 5.2 reviews background on CT reconstruction and sequential optimal experimental design using RL. Section 5.3 presents the proposed reconstruction and scanning framework. Section 5.4 reports experimental results, and Sections 5.5–5.6 provide discussion and conclusions.

5.2 Background

5.2.1 CT Reconstruction

In X-ray CT, the objective is to reconstruct an unknown image $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, representing the spatial distribution of attenuation coefficients within an object, from a set of external measurements. Let Θ denote the full set of available candidate projection angles. At step k, the agent selects one angle $\theta_k \in \Theta$. The sequence of selected angles is denoted by $\boldsymbol{\theta}$. These measurements, denoted $\boldsymbol{y}(\boldsymbol{\theta})$, are acquired at a sequence of design parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$, typically projection angles. The image formation process is described by a forward operator $\boldsymbol{A}(\boldsymbol{\theta})$, which discretizes the Radon transform under the specified acquisition geometry [38].

To account for measurement uncertainty, we adopt a statistical observation model in which the measurements are treated as random variables:

$$\boldsymbol{y}(\boldsymbol{\theta}) \sim \pi_{\text{data}}(\boldsymbol{y} \mid \bar{\boldsymbol{x}}; \boldsymbol{\theta}), \tag{5.1}$$

where π_{data} denotes the data likelihood.

Beer–Lambert law. X-ray attenuation follows the Beer–Lambert law [38], which relates transmitted intensity to the cumulative attenuation along an X-ray path. Let $I_0(\theta_j)$ denote the incident photon flux (dose) at angle θ_j, and $I_i(\theta_j)$ the transmitted intensity measured at detector bin i. Then the attenuation model is

$$I_i(\theta_j) = I_0(\theta_j) \exp(-[\boldsymbol{A}(\theta_j)\bar{\boldsymbol{x}}]_i)\,, \tag{5.2}$$

where $[\boldsymbol{A}(\theta_j)\bar{\boldsymbol{x}}]_i$ represents the line integral of $\bar{\boldsymbol{x}}$ along ray i at angle θ_j. The corresponding noise-free measurement is

$$y_i(\theta_j) = -\log\left(\frac{I_i(\theta_j)}{I_0(\theta_j)}\right) = [\boldsymbol{A}(\theta_j)\bar{\boldsymbol{x}}]_i. \tag{5.3}$$

In practice, the observed measurements $\boldsymbol{y}(\theta_j)$ deviate from this ideal due to stochastic noise governed by the dose level $I_0(\theta_j)$.

Bayesian inverse formulation. Reconstructing $\bar{\boldsymbol{x}}$ from noisy measurements is an ill-posed inverse problem. Within a Bayesian framework, the posterior distribution

combines the measurement model with prior information:

$$\begin{aligned}\widehat{\boldsymbol{x}}(\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{x}} \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{x}} \frac{\pi_{\text{data}}(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) \cdot \pi_{\text{prior}}(\boldsymbol{x})}{\pi(\boldsymbol{y})},\end{aligned} \tag{5.4}$$

where $\pi_{\text{prior}}(\boldsymbol{x})$ encodes assumptions about the image, such as smoothness or sparsity. The maximum a posteriori (MAP) estimate thus balances data fidelity and prior knowledge to yield a stable reconstruction.

5.2.2 Sequential Optimal Experimental Design via Reinforcement Learning

Reinforcement learning (RL) provides a natural framework for solving sequential optimal experimental design (OED) problems in CT imaging by modeling the acquisition process as a *Partially Observable Markov Decision Process (POMDP)* [77]. This perspective has been effectively applied in prior work [97, 96, 99], where the agent interacts with the CT environment to iteratively select measurements and refine the reconstruction.

The interaction starts from an initial estimate $\widehat{\boldsymbol{x}}_1$, typically a zero image. At each time step k, the agent selects an acquisition angle θ_k based on the current reconstruction $\widehat{\boldsymbol{x}}_k$. A noisy measurement $\boldsymbol{y}_k$ is then acquired using the forward model in Equation (5.1), and a new reconstruction $\widehat{\boldsymbol{x}}_{k+1}$ is computed based on all measurements up to step k. A reward r_k is subsequently calculated by comparing $\widehat{\boldsymbol{x}}_{k+1}$ to the ground truth image.

The agent learns a parameterized policy $\pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$, which maps the current reconstruction to a probability distribution over candidate angles. After M steps, the acquisition process produces a trajectory $\boldsymbol{\tau}$ defined as:

$$\boldsymbol{\tau} = \{\widehat{\boldsymbol{x}}_1, (\theta_1, \widehat{\boldsymbol{x}}_2, r_1), \ldots, (\theta_M, \widehat{\boldsymbol{x}}_{M+1}, r_M)\}.$$

This trajectory follows a distribution governed by:

$$\pi_{\text{chain}}(\boldsymbol{\tau}; \boldsymbol{w}_a) = \prod_{k=1}^{M} \pi_{\text{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \cdot \pi_t(\widehat{\boldsymbol{x}}_{k+1} \mid \widehat{\boldsymbol{x}}_k, \theta_k), \tag{5.5}$$

where $\pi_t(\widehat{\boldsymbol{x}}_{k+1} \mid \widehat{\boldsymbol{x}}_k, \theta_k)$ represents the transition model capturing the effect of measurement and reconstruction.

The learning objective is to maximize the expected cumulative reward, where future rewards may be discounted using a factor $\gamma \in [0, 1]$ to prioritize immediate outcomes. The state-value function under policy π_{a} is defined as:

$$V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_k) = \mathbb{E}_{\boldsymbol{\tau} \sim \pi_{\text{chain}}} \left[\sum_{t=k}^{M} \gamma^{\,t-k} r_t \,\middle|\, \widehat{\boldsymbol{x}}_k \right], \tag{5.6}$$

where γ^{t-k} discounts the reward obtained at time step k relative to the current step k. This formulation captures the preference for receiving rewards earlier rather than later.

The overall objective in RL is to maximize the expected return starting from the fixed initial zero image $\widehat{\boldsymbol{x}}_1$::

$$J(\boldsymbol{w}_a) = V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_1), \tag{5.7}$$

where optimization is performed over the policy parameters $\boldsymbol{w}_a$.

5.2.3 Policy Gradient Methods

Policy gradient methods aim to optimize the expected return $J(\boldsymbol{w}_a)$ by computing gradients with respect to the policy parameters $\boldsymbol{w}_a$. A popular approach is the *actor-critic* framework, where the *actor* represents the policy, and the *critic* estimates the state-value function to guide policy improvement.

To reduce variance and improve learning efficiency, policy updates are often scaled by an *advantage estimate* $\widehat{A}_k$, which measures how much better (or worse) an action is compared to the critic's baseline value estimate $V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_k)$. A simple and widely used choice is the one-step Temporal-Difference (TD(0)) estimator [84]:

$$\widehat{A}_k = r_k + \gamma V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_{k+1}) - V^{\pi_{\text{a}}}(\widehat{\boldsymbol{x}}_k), \tag{5.8}$$

where γ is the discount factor. More general estimators, such as Generalized Advantage Estimation (GAE) [71], can further reduce variance by incorporating multi-step returns.

The policy gradient can then be approximated from sampled trajectories as:

$$\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a) \approx \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{k=1}^{M} \nabla_{\boldsymbol{w}_a} \log \pi_{\text{a}}(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a) \cdot \widehat{A}_k \right), \tag{5.9}$$

where N is the number of sampled trajectories and M is the number of timesteps per

trajectory.

5.3 Methods

This section introduces the proposed framework, which combines a dose-aware reconstruction algorithm with prior information to handle unequal dose allocation, and a RL approach to address the adaptive CT scanning problem.

5.3.1 Reconstruction with Unequal Dose Allocation

Building on the Beer–Lambert formulation in Section 5.2, we now incorporate measurement noise arising from photon statistics. Since photon detection follows a Poisson distribution, the measured photon count at detector bin i for angle θ_j can be modeled as (according to Equation (5.2))

$$I_i(\theta_j) \sim \text{Poisson}(I_0(\theta_j)\, \exp(-[\boldsymbol{A}(\theta_j)\bar{\boldsymbol{x}}]_i))\,. \tag{5.10}$$

where $\boldsymbol{I}_0(\theta_j) = [I_{0,1}(\theta_j), \ldots, I_{0,n_{\text{det}}}(\theta_j)]^T$ denotes the incident photon flux at all detector bins. (For the experiments in this paper, we consider a spatially uniform incident dose so that $\boldsymbol{I}_0(\theta_j) = I_0(\theta_j)\mathbf{1}$.)

Normalizing by $\boldsymbol{I}_0(\theta_j)$ and applying a negative logarithm yields the noisy log-transformed measurements:

$$\boldsymbol{y}(\theta_j) = -\log\Big(\tfrac{\boldsymbol{I}(\theta_j)}{\boldsymbol{I}_0(\theta_j)}\Big)\,. \tag{5.11}$$

From the Bayesian inverse formulation in Equation (5.4), the negative log-likelihood (up to constants) gives the dose-aware reconstruction objective:

$$\begin{aligned}\widehat{\boldsymbol{x}}(\boldsymbol{\theta}) \propto \arg\min_{\boldsymbol{x}} \Big\{ &\sum_j \boldsymbol{I}_0(\theta_j)^T \exp(-\boldsymbol{A}(\theta_j)\boldsymbol{x}) \\ &+ \sum_j \big[\boldsymbol{I}_0(\theta_j) \odot \exp\big(-\boldsymbol{y}(\theta_j)\big)\big]^T \boldsymbol{A}(\theta_j)\boldsymbol{x} - \log \pi_{\text{prior}}(\boldsymbol{x})\Big\}.\end{aligned} \tag{5.12}$$

This formulation explicitly incorporates unequal per-angle dose levels $I_0(\theta_j)$, which directly affect both the noise characteristics of the measurements and their contribution to the reconstruction objective.

PWLS formulation. Direct minimization of Equation (5.12) is computationally demanding. Using a second-order Taylor expansion around the measured log-data [38] yields a Penalized Weighted Least-Squares (PWLS) problem [62, 95, 38]:

$$\widehat{\boldsymbol{x}}(\boldsymbol{\theta}) \propto \arg\min_{\boldsymbol{x}} \left\{ \tfrac{1}{2} (\boldsymbol{y}(\boldsymbol{\theta}) - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x})^T \boldsymbol{W}(\boldsymbol{\theta}) (\boldsymbol{y}(\boldsymbol{\theta}) - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}) - \log \pi_{\text{prior}}(\boldsymbol{x}) \right\}, \quad (5.13)$$

with diagonal weight matrix

$$\boldsymbol{W}(\boldsymbol{\theta}) = \operatorname{diag}\Big(\boldsymbol{I}_0(\boldsymbol{\theta}) \odot \exp\big(-\boldsymbol{y}(\boldsymbol{\theta}) \big) \Big). \quad (5.14)$$

Here, $\boldsymbol{I}_0(\boldsymbol{\theta})$ expands the per-angle dose allocation $I_0(\theta_j)$ to all bins. In contrast to conventional PWLS formulations [95] that assume a uniform incident dose across views, this model explicitly accounts for angle-dependent dose allocation. This *dose-aware* weighting provides a principled mechanism to assign higher statistical confidence to high-dose projections and lower confidence to noisy low-dose measurements, effectively capturing the signal-dependent variance characteristic of Poisson noise in X-ray CT.

Plug-and-Play priors In particular when using low radiation dosages distributed over few projection angles, minimizing the PWLS objective directly is ill-posed and leads to very noisy reconstructions ridden with sparse-angle artifacts. To regularize the inversion in the most flexible way, we describe here how to employ the Plug-and-Play (PnP) framework [93]. PnP decouples the reconstruction into two alternating steps: a gradient step for the PWLS objective followed by the application of an image denoiser. To illustrate this, we implemented denoising with a Total Variation (TV) energy [65, 20], denoted by $\text{prox}_{\tau\,\text{TV}}(\cdot)$. TV penalizes the ℓ_1-norm of image gradients and is used to suppresses noise in homogeneous regions while preserving sharp edges. While this choice means that the PnP framework reduces to a proximal gradient descent with an explicit image prior $\pi_{\text{prior}}(\boldsymbol{x})$, using, e.g., a trained DNN would correspond to using an implicit prior.

Explanation. Algorithm 4 performs gradient descent on the PWLS objective with explicit modeling of angle-dependent dose allocation. At each iteration, the current reconstruction is forward-projected and compared with the measured log-projections to form a residual. This residual is reweighted by a Poisson–dose factor, which increases the influence of high-dose views and suppresses noisy low-dose views. The weighted residual is then backprojected to compute the gradient, and the step size is normalized by the estimated Lipschitz constant (Appendix C.1) to ensure stable convergence. After the gradient update, a TV denoiser $\text{prox}_{\tau\,\text{TV}}(\cdot)$ is applied in the PnP framework,

Algorithm 4 Dose-Aware PnP–PWLS Reconstruction (Gradient + TV Prior)

Require: Forward operator $\boldsymbol{A}(\boldsymbol{\theta})$; measured log-projections $\boldsymbol{y}(\boldsymbol{\theta})$; dose allocation $\boldsymbol{d}(\boldsymbol{\theta})$; number of iterations T; initial estimate $\boldsymbol{x}_0$; PnP denoiser strength $\tau > 0$ (TV parameter); positivity flag.

Ensure: Reconstructed image $\boldsymbol{x}_T$.

$\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
Normalize dose: $\boldsymbol{d}(\boldsymbol{\theta}) \leftarrow \boldsymbol{d}(\boldsymbol{\theta})/\mathrm{mean}(\boldsymbol{d}(\boldsymbol{\theta}))$
Dose–Poisson weights: $\boldsymbol{R}(\boldsymbol{\theta}) \leftarrow \boldsymbol{d}(\boldsymbol{\theta}) \odot \exp\big(-\boldsymbol{y}(\boldsymbol{\theta})\big)$
Define weighting operator: $\mathcal{W}(\cdot) : \boldsymbol{u} \mapsto \boldsymbol{R}(\boldsymbol{\theta}) \odot \boldsymbol{u}$
Estimate Lipschitz constant L via Appendix C.1; set step size $\alpha \leftarrow h/L$ with $h \in [0, 2]$
for $k = 1$ **to** T **do** ▷ Data fidelity + PnP prior
  $\boldsymbol{y}_{\mathrm{pred}} \leftarrow \boldsymbol{A}(\boldsymbol{\theta})\,\boldsymbol{x}$ ▷ Forward projection
  $\boldsymbol{r} \leftarrow \mathcal{W}(\boldsymbol{y}_{\mathrm{pred}} - \boldsymbol{y}(\boldsymbol{\theta}))$ ▷ Dose-aware residual
  $\boldsymbol{g} \leftarrow \boldsymbol{A}(\boldsymbol{\theta})^{\top}\boldsymbol{r}$ ▷ Backprojection (gradient of data term)
  $\boldsymbol{x} \leftarrow \boldsymbol{x} - \alpha\,\boldsymbol{g}$ ▷ Gradient step (data fidelity)
  $\boldsymbol{x} \leftarrow \mathrm{prox}_{\tau\,\mathrm{TV}}(\boldsymbol{x})$ ▷ PnP denoiser: TV proximal operator (e.g. Chambolle's algorithm)
  if positivity flag **then**
    $\boldsymbol{x} \leftarrow \max(\boldsymbol{x}, 0)$
  end if
end for
return $\boldsymbol{x}$

providing noise suppression while preserving edges. Finally, a positivity constraint can be optionally enforced to maintain physically valid attenuation values.

Compared with standard iterative reconstruction, this dose-aware PnP scheme explicitly accounts for non-uniform photon allocation across projection angles, yielding statistically consistent updates under unequal-dose settings while benefiting from edge-preserving regularization.

5.3.2 Proximal Policy Optimization for CT Environment

To stably optimize the objective $J(\boldsymbol{w}_a)$ in Equation (5.7), Proximal Policy Optimization (PPO) [72] is adopted as the RL framework. The vanilla policy-gradient method in Equation (5.9) updates the policy parameters by directly ascending an unbiased stochastic estimate of $\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a)$, computed from sampled trajectories, without explicitly constraining the size of policy updates. PPO replaces this unconstrained update with a clipped surrogate objective that approximates the same gradient while preventing excessively large policy changes, thereby improving training stability:

$$Loss^{\text{actor}}(\boldsymbol{w}_a) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{M}\min\Big(e_k(\boldsymbol{w}_a)\,\widehat{A}_k,\; c_{1-\epsilon}^{1+\epsilon}\big(e_k(\boldsymbol{w}_a)\big)\,\widehat{A}_k\Big), \tag{5.15}$$

where $e_k(\boldsymbol{w}_a) = \frac{\pi_a(\theta_k|\widehat{\boldsymbol{x}}_k;\boldsymbol{w}_a)}{\pi_a^{\text{old}}(\theta_k|\widehat{\boldsymbol{x}}_k)}$ is the probability ratio between new and old policies, $\widehat{A}_k$ is the estimated advantage, and $\epsilon > 0$ is a clipping parameter. Here, $c_{1-\epsilon}^{1+\epsilon}(\cdot)$ clips the probability ratio $e_k(\boldsymbol{w}_a)$ to the interval $[1-\epsilon, 1+\epsilon]$, ensuring stable and conservative policy updates. The objective is maximized during training to update the actor parameters $\boldsymbol{w}_a$.

The critic is trained using a mean squared error (MSE) loss with respect to the Monte Carlo return G_k (Formally, $G_k = \sum_{t=k}^{M}\gamma^{t-k}r_t$.):

$$Loss^{\text{critic}}(\boldsymbol{w}_v) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{M}\big(G_k - V(\widehat{\boldsymbol{x}}_k;\boldsymbol{w}_v)\big)^2, \tag{5.16}$$

Reward Definition. The reward function depends on the objective of the reconstruction task. For general image quality assessment, we use the Peak Signal-to-Noise Ratio (PSNR), computed over a region of interest (RoI) Ω:

$$\text{PSNR}_\Omega(\bar{\boldsymbol{x}}, \widehat{\boldsymbol{x}}) = 10\log_{10}\left(\frac{H^2}{\frac{1}{|\Omega|}\sum_{i\in\Omega}(\bar{x}_i - \widehat{x}_i)^2}\right), \tag{5.17}$$

where H denotes the dynamic range of the image.

For defect-focused tasks, the reward can instead be based on the Contrast-to-Noise Ratio (CNR):

$$\mathrm{CNR}_{\mathcal{D},\mathcal{B}}(\bar{\boldsymbol{x}}, \widehat{\boldsymbol{x}}) = \frac{\mu_{\mathcal{D}} - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}}, \tag{5.18}$$

where $\mu_{\mathcal{D}}$ is the mean intensity within the defect mask $\mathcal{D}$, and $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ denote the mean and standard deviation in a background region $\mathcal{B}$.

Training Procedure. The RL agent interacts with the CT environment under a fixed dose budget B, which is spent in base quanta Δb per projection. Dose allocation is modeled implicitly by repeating angle selections: if an angle is chosen z times, the allocated dose is $z\Delta b$. The episode terminates after $L = \lfloor B/\Delta b \rfloor$ steps, corresponding to exhaustion of the total dose budget B. The policy maps the current reconstruction $\widehat{\boldsymbol{x}}_k$ to a probability distribution over candidate projection angles, thereby jointly determining angle selection and dose allocation. The overall training loop is summarized in Algorithm 5. A trajectory buffer stores observations, actions, rewards, and log-probabilities of the old policy. After each rollout, advantages and returns are computed, and the policy and value networks are updated by Equations (5.15)–(5.16). Implementation details including advantage estimation, entropy regularization, and minibatch updates are provided in Appendix C.2.

Intuition. The PPO framework enables the agent to learn dose-aware acquisition strategies directly through interaction with the CT environment. Rather than distributing photons uniformly across all projection angles, the learned policy adaptively selects which angles to measure at each step. Since a fixed dose Δb is used for every acquisition, repeatedly selecting the same angle results in a higher cumulative dose being allocated to that direction, while angles that are selected fewer times receive less total dose. In this way, the policy implicitly concentrates the dose budget on more informative views through repeated measurements, without reallocating or reducing the dose that has already been acquired. The critic provides value estimates that stabilize training, and the reward function drives the policy to optimize either overall reconstruction quality (PSNR) or defect detectability (CNR). This design balances exploration of new angles with exploitation of informative ones, leading to improved reconstruction quality under limited-dose conditions.

The overall pipeline is illustrated in Figure (5.1). At each step k, the dose-aware reconstruction produces the current state $\widehat{\boldsymbol{x}}_k$. The PPO agent then selects an angle θ_k based on the current policy π_k, and the environment returns the next state $\widehat{\boldsymbol{x}}_{k+1}$

Algorithm 5 PPO Training in a CT Environment with a Dose Budget

Require: Environment $\mathcal{E}$, policy $\pi_a(\cdot; \boldsymbol{w}_a)$, value function $V(\cdot; \boldsymbol{w}_v)$, total dose budget B, fixed dose per step Δb, number of trajectories per iteration n

Set trajectory length $L \leftarrow \left\lfloor \frac{B}{\Delta b} \right\rfloor$
Set horizon $T \leftarrow n \cdot L$ ▷ total steps per PPO iteration
for each iteration **do**
    Initialize empty trajectory buffer $\mathcal{T}$
    for $i = 1$ **to** n **do** ▷ collect n trajectories
        Reset environment: $\widehat{\boldsymbol{x}}_1$(zero matrix) $\leftarrow \mathcal{E}$.reset()
        for $k = 1$ **to** L **do**
            Sample angle $\theta_k \sim \pi_a(\cdot \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$
            $\log \pi_k \leftarrow \log \pi_a(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$, $v_k \leftarrow V(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$
            Apply action: $(\widehat{\boldsymbol{x}}_{k+1}, r_k) \leftarrow \mathcal{E}$.step$(\theta_k)$
            $d_k \leftarrow \mathbb{I}[k = L]$
            Store $(\widehat{\boldsymbol{x}}_k, \theta_k, r_k, \widehat{\boldsymbol{x}}_{k+1}, \log \pi_k, v_k, d_k)$ in $\mathcal{T}$
        end for
    end for
    Compute advantages $\widehat{A}_k$ and returns G_k from $\mathcal{T}$ (Appendix C.2)
    Normalize advantages
    Update actor by maximizing the clipped PPO objective (Equation (5.15)), where π^{old} uses stored $\log \pi_k$ from $\mathcal{T}$
    Update critic by minimizing the MSE loss (Eq. (5.16))
end for

along with the reward r_k. The transition $(\widehat{\boldsymbol{x}}_k, \theta_k, r_k, \widehat{\boldsymbol{x}}_{k+1}, \log \pi_k, v_k, d_k)$ is appended to the buffer $\mathcal{T}$ for PPO updates. The pipeline continues until the dose budget is exhausted. Orange blocks and dashed arrows denote components used only during training, whereas in real applications only the blue blocks and solid arrows are executed.

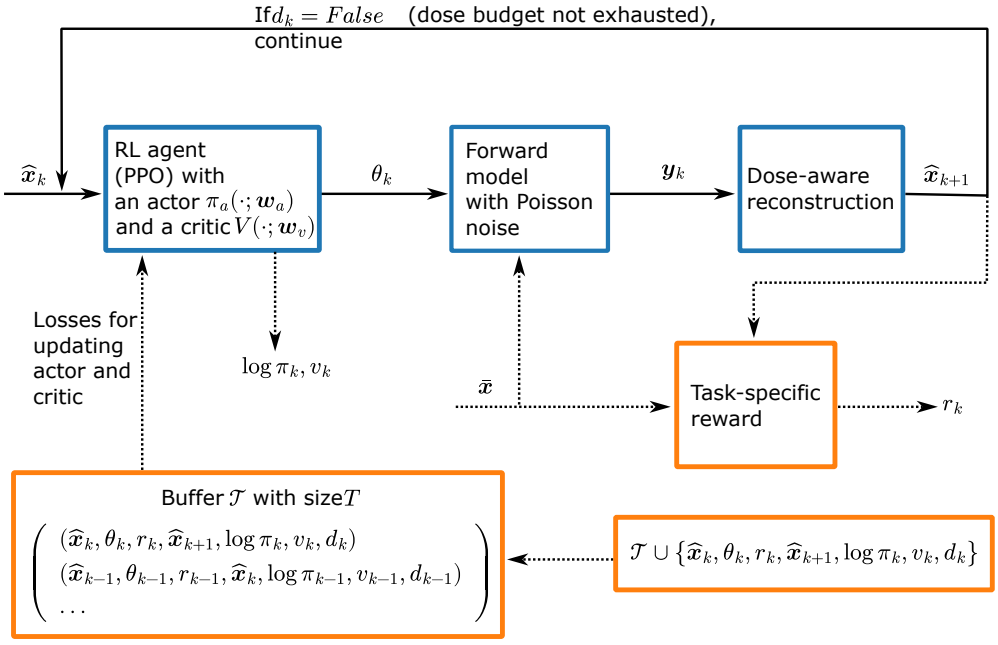


Figure 5.1: Pipeline of the proposed method. At each step k, the dose-aware reconstruction produces the current state $\widehat{\boldsymbol{x}}_k$. The PPO agent selects an angle θ_k based on the current policy π_k, and the environment returns the next state $\widehat{\boldsymbol{x}}_{k+1}$ and reward r_k. The transition $(\widehat{\boldsymbol{x}}_k, \theta_k, r_k, \widehat{\boldsymbol{x}}_{k+1}, \log \pi_k, v_k, d_k)$ is stored in the buffer $\mathcal{T}$ for PPO updates. The pipeline continues until the dose budget is exhausted. Orange blocks and dashed arrows denote components used only during training, whereas in real applications only the blue blocks and solid arrows are executed.

5.4 Results

We present experimental results to validate both components of the proposed framework. First, we evaluate the dose-aware PnP–PWLS reconstruction algorithm against standard iterative methods. Next, we assess the RL policy for joint angle selection and dose allocation, using the golden ratio policy as a baseline. Finally, we demonstrate the effectiveness of the proposed approach on a specific task of interest—defect

detection.

5.4.1 Datasets

In our numerical experiments, we consider several phantoms illustrated in Figure (5.2), each with a size of 256×256 pixels. The phantoms are composed of multiple materials and grouped into three datasets. Dataset (a) contains wedge-shaped phantoms composed of different materials, with variations in rotation, position, and size. These phantoms are inspired by industrial CT inspection scenarios involving fiber-reinforced or layered composite materials, where internal structures exhibit strong directional characteristics. Datasets (b) and (c) consist of ellipse- and foam-based phantoms [61] with different material properties. These phantoms are inspired by polymer or metal foam–like materials that act as containers for embedded foreign objects or defects with preferential orientations. Such configurations are representative of industrial CT inspection tasks, where the goal is to detect and characterize internal objects or anisotropic defects within a heterogeneous, porous background. In particular, dataset (c) includes ellipse-based phantoms containing defects. All datasets include phantoms at different rotations, represented by 36 equally spaced angles between $0°$ and $179°$. In addition, the phantoms exhibit variations in scaling and spatial shifts.

5.4.2 Implementation

Tomosipo [40] was used to generate simulated projections and perform reconstructions. The image size was set to 256×256, and a projection size of 384 was chosen to ensure that each view fully covered the image domain.

The photon budgets in our synthetic experiments are normalized photon-count parameters in the Poisson noise model and are not directly calibrated to physical dose units such as mGy. Such a conversion would require scanner-specific information, including the X-ray spectrum, filtration, exposure time, detector efficiency, acquisition geometry, and object attenuation. We therefore use these budgets to define controlled low-count regimes for comparing angle and dose selection policies under matched relative dose constraints. They should be interpreted as representative of dose-scarce imaging scenarios, such as low-dose CT, repeated or interventional imaging, and industrial CT inspection with limited exposure or acquisition time, rather than as a one-to-one mapping to a specific clinical or industrial dose value.

For the reconstruction algorithm, the parameter h in Algorithm 4 was set to 1.8, and the PnP–TV denoiser employed a fixed proximal step size of 0.0005. In the RL

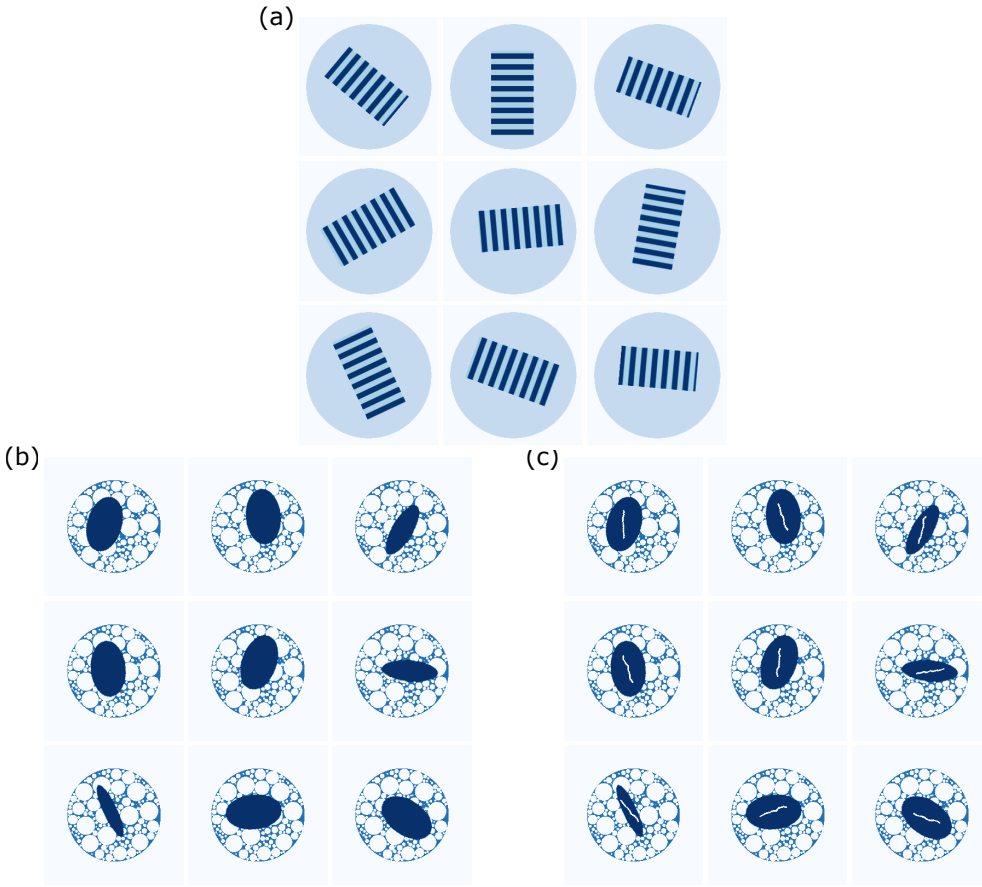


Figure 5.2: Example training dataset consisting of three types of phantoms. (a) Wedge-shaped phantoms composed of two different materials, with variations in rotation, position, and size. (b) Ellipse- and foam-based phantoms with different material properties. (c) Ellipse-based phantoms containing defects. The ellipses vary in rotation, position, size, and roundness, providing a diverse set of training samples.

framework (Algorithm 5), the discount factor was set to $\gamma = 0.99$, and the replay buffer capacity was fixed at 500 transitions. The actor and critic losses were weighted by 1.0 and 0.5, respectively, and an entropy regularization term with weight 0.01 was included to promote adequate exploration during training (see Appendix C.2 for details). These hyperparameters were selected empirically to provide a stable balance between policy optimization, accurate value-function estimation, and robust exploratory behavior. A convolutional encoder was employed to extract low-dimensional feature representations from the input images, following the design in [97]. All network parameters were optimized using the Adam algorithm [47] with a learning rate of 10^{-4} and a weight decay of 10^{-5}. Further architectural and algorithmic details are provided in Table (C.1) in Appendix C.2.

5.4.3 Experiment 1: Reconstruction Using Dose-Aware PnP–PWLS with a TV Prior

This experiment evaluates the proposed reconstruction algorithm against standard iterative methods. We focus on datasets (a) and (b) described in Section 5.4.1. Reconstructions were performed with three equal angular sampling settings: 36, 48, and 60 angles. To simulate unequal dose allocation across views, odd-indexed angles were assigned a low dose (100 photons), and even-indexed angles a high dose (1000 photons). The number of gradient descent iterations was varied from 5 to 100 in increments of 5.

Four methods were compared: Simultaneous Iterative Reconstruction Technique (SIRT), conventional PWLS, dose-aware PWLS, and dose-aware PnP–PWLS with a TV prior. All reconstructions were performed under a non-negativity constraint.

Figure (5.3) compares reconstruction performance on datasets (a) and (b). Both dose-aware variants (PWLS and PnP–PWLS with a TV prior) consistently achieve higher PSNR and exhibit greater robustness to the number of iterations across different angular sampling settings. In contrast, SIRT and conventional PWLS degrade noticeably at higher iteration counts, making reconstruction quality more sensitive to the iteration choice and less reliable for subsequent experiments. These results demonstrate that dose-aware weighting improves reconstruction stability, with PnP–PWLS with a TV prior consistently providing the best overall performance.

Remark. The performance gap between the proposed dose-aware methods and the baselines becomes more pronounced at higher iteration counts. While SIRT and conventional PWLS tend to overfit the noise and degrade in quality, dose-aware weight-

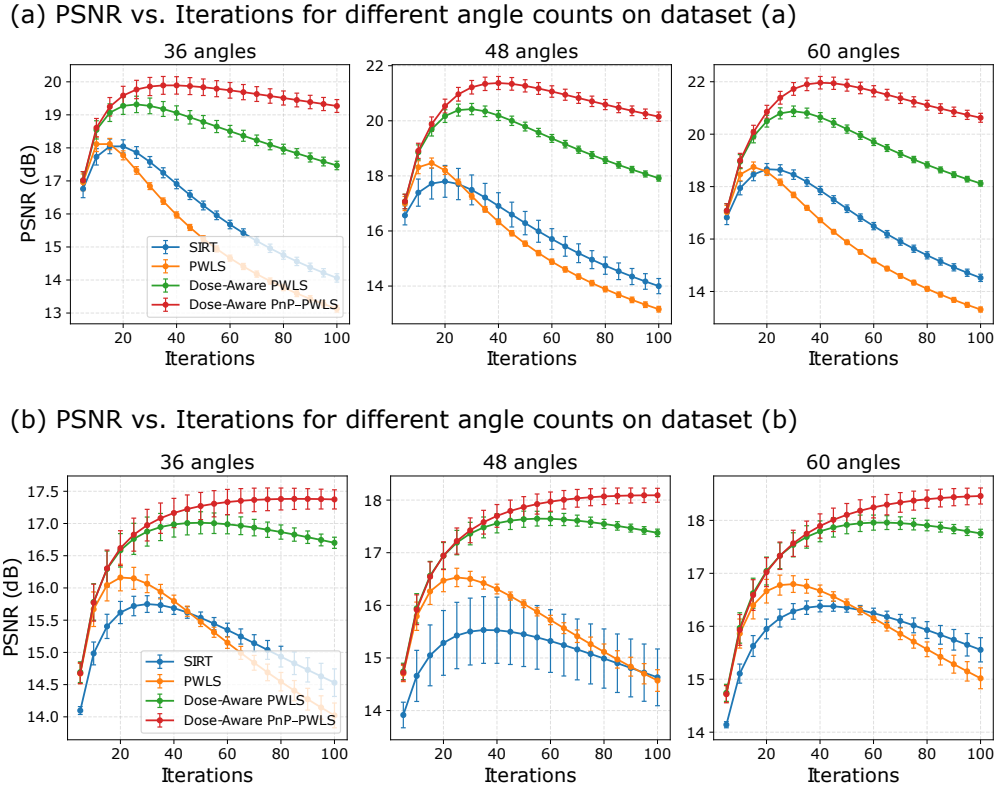


Figure 5.3: PSNR versus number of iterations for different angular sampling settings (36, 48, and 60 angles, left to right) on datasets (a) and (b). The four methods compared are SIRT, conventional PWLS, dose-aware PWLS, and dose-aware PnP–PWLS. Curves show the mean PSNR across all non-defective phantoms in the dataset, and error bars indicate the standard deviation.

ing preserves relative stability by scaling the contribution of each view according to its photon statistics, although some degradation remains visible on dataset (a). This highlights that dose-awareness is particularly beneficial in long-iteration regimes, where conventional methods become unreliable. In addition, incorporating a TV prior provides effective denoising while preserving edges, further improving reconstruction quality.

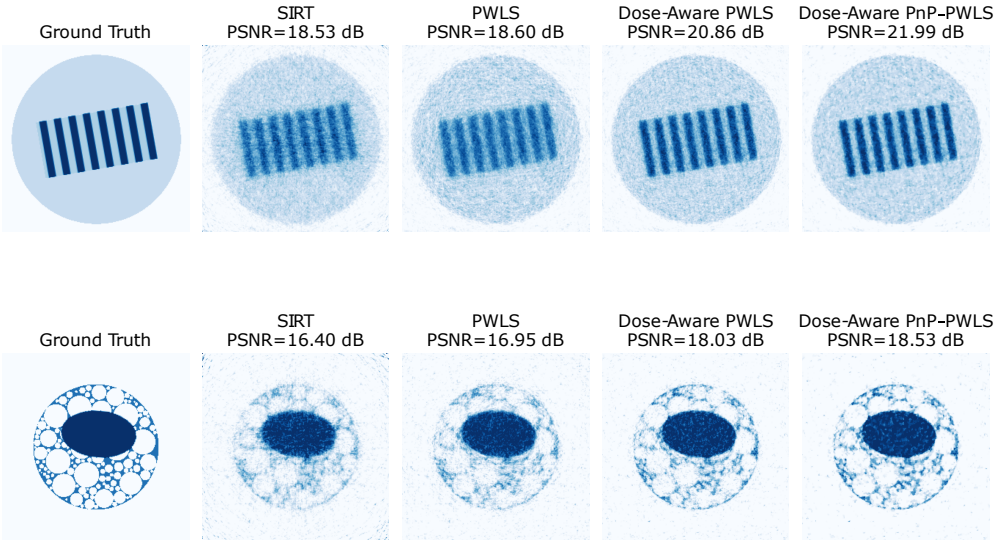


Figure 5.4: Reconstruction results for two phantoms with 60 projection angles using SIRT, conventional PWLS, and the proposed dose-aware methods. For each method, the iteration count was selected from 0 to 100 to yield its best reconstruction. SIRT reconstructions suffer from low contrast and artifacts, while PWLS provides modest improvements but remains noisy. The proposed dose-aware approaches achieve the highest PSNR and visibly improve contrast.

These visual results are consistent with the quantitative PSNR values reported in Figure (5.4), where the proposed dose-aware methods achieve improvements of approximately 1.5–3 dB over SIRT and conventional PWLS when considering their best reconstructions within the iteration range of 0–100. This demonstrates that explicitly modeling angle-dependent dose allocation not only enhances numerical performance but also leads to visibly cleaner and higher-contrast reconstructions.

Experiment 2: Angle Selection and Dose Allocation for Reconstruction Quality Improvement

This experiment evaluates whether the RL agent can jointly select projection angles and allocate dose to improve reconstruction quality when the reward function is based solely on PSNR. We used dataset (a) for this study and adopted the dose-aware PnP–PWLS reconstruction method with a TV prior, which showed the best performance in the previous experiments.

We compared the learned RL policy with two baselines. The first baseline is the *Golden Ratio (GR) policy* [49, 21], which is commonly used in sparse-view CT. The GR policy generates projection angles by applying an irrational angular increment, resulting in a non-repeating sequence that approximately uniformly fills the angular range over time. The second baseline is a *Projected-Uncertainty (PU) adaptive policy*, which provides a simple adaptive non-RL comparison.

For the PU baseline, an uncertainty map is computed at acquisition step k from the change between consecutive reconstructions,

$$u_{k-1} = |\widehat{\boldsymbol{x}}_{k-1} - \widehat{\boldsymbol{x}}_{k-2}|\,. \tag{5.19}$$

Each candidate angle $\theta \in \Theta$ is scored by forward-projecting this uncertainty map,

$$S_k(\theta) = \frac{\|\boldsymbol{A}(\theta)u_{k-1}\|_1}{1 + n_k(\theta)}, \tag{5.20}$$

where $\boldsymbol{A}(\theta)$ is the single-view forward operator at angle θ, and $n_k(\theta)$ is the number of previous selections of angle θ. The denominator discourages repeated selection of the same angle while still allowing additional dose to be assigned through repeated measurements. The next angle is selected greedily as

$$\theta_k = \arg\max_{\theta\in\Theta} S_k(\theta). \tag{5.21}$$

For the first acquisition step, the first GR angle is used to avoid the degenerate case of uniform initial uncertainty. This baseline uses the same candidate angle set, dose budget, repeated-angle dose-allocation mechanism, Poisson noise model, and dose-aware PnP–PWLS reconstruction method as the RL policy, but it is myopic and does not learn a task-specific acquisition strategy.

Training procedure. The agent was trained on 150 phantoms from dataset (a). For the GR policy, 20 angles were used with an equal dose assigned to each angle. Three separate RL policies were trained for three different total photon budgets of 6000 photons (300 photons per angle), 7000 photons (350 photons per angle) and 8000 photons (400 photons per angle). For the RL policy, the agent could allocate more dose to informative angles by repeating them. For example, with a base dose of 300 photons per angle, if angle a is repeated n times, the total dose assigned to that angle becomes $n \times 300$. Training was performed for 1500 episodes using the reward function in Eq. (5.22), assuming that the ground truth $\bar{\boldsymbol{x}}$ was non-defective. The weight parameter α in Eq. (5.22) was set to 0.2, and the RoI Ω corresponded to the full image.

Validation procedure. Validation was performed on a separate test set of 100 phantoms with unseen rotation angles. All results reported in this section are obtained from this validation set using the trained policy.

The reward functions used in this work require ground-truth images and are therefore used only during offline training in simulation. During inference, the trained policy does not require access to the ground truth; it selects angles and allocates dose based only on the current reconstruction state and acquisition history.

Quantitative comparison. Table 5.1 summarizes the performance of the GR, PU, and RL policies on the unseen-rotation test set across total photon budgets of 6000, 7000, and 8000. The RL policy consistently achieves higher PSNR than both the non-adaptive GR policy and the adaptive PU baseline. Compared with GR, the mean PSNR improvements of the RL policy are 2.80, 2.46, and 2.68 dB for total photon budgets of 6000, 7000, and 8000, respectively. The corresponding 95% paired bootstrap confidence intervals are strictly positive, and the improvements remain significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$. In contrast, the PU baseline does not outperform GR, indicating that a simple myopic projected-uncertainty strategy is insufficient for this wedge reconstruction task. As the total photon budget increases, the RL policy selects a larger number of distinct angles, suggesting that repeated angle selection is reduced when more dose is available.

*Significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$.

Angle selection and reconstruction analysis. Figures (5.5) and (5.6) (a1–a2) illustrate the angle selections produced by the GR and RL policies for wedge-shaped

Table 5.1: Performance comparison of GR, projected-uncertainty (PU), and RL policies on unseen-rotation tests for the wedge phantom experiment. Results are reported as mean ± standard deviation. The RL–GR difference is reported with 95% paired bootstrap confidence intervals. One-sided p-values for RL outperforming GR were corrected using Benjamini–Hochberg FDR control at $\alpha = 0.05$.

Total photons	Policy	PSNR	Average number of angles	PSNR diff. RL–GR [95% CI]
6000	GR	16.82 ± 0.40	20	2.80 [2.76, 2.83]*
	PU	15.64 ± 1.15	20	
	RL	19.61 ± 0.67	7.26 ± 1.13	
7000	GR	17.25 ± 0.43	20	2.46 [2.42, 2.50]*
	PU	15.92 ± 1.23	20	
	RL	19.71 ± 0.89	7.92 ± 1.44	
8000	GR	17.60 ± 0.47	20	2.68 [2.65, 2.71]*
	PU	16.15 ± 1.30	20	
	RL	20.28 ± 0.63	8.81 ± 1.32	

phantoms at total doses of 6000 and 8000 photons, respectively. Each colored spoke denotes the beam direction of a parallel-beam view at angle θ (degrees), measured counter-clockwise from the x-axis, with θ and $\theta + 180^\circ$ being equivalent. The color coding indicates the number of repeats, corresponding to the relative photon allocation at that angle. The GR policy distributes views uniformly across the angular range, whereas the RL policy adapts its strategy by concentrating angles and dose around preferential directions corresponding to the wedge orientations. This adaptive allocation provides more informative measurements along high-contrast directions while reducing sampling in less informative regions.

Figures (5.5) and (5.6) (b1–b4) present reconstruction results for wedge-shaped phantoms using the GR and RL policies at total doses of 6000 and 8000 photons. For both dose levels, reconstructions obtained with the GR policy appear noisy and exhibit poor contrast between materials. In contrast, the RL policy produces reconstructions with clearer separation of material regions and reduced noise, demonstrating the advantage of adaptive angle selection and dose allocation.

Experiment 3: Angle Selection and Dose Allocation for Defect Detectability Improvement

This experiment addresses the task of defect detection. Both dataset (b) (non-defective) and dataset (c) (defective) phantoms were used to evaluate whether the agent can

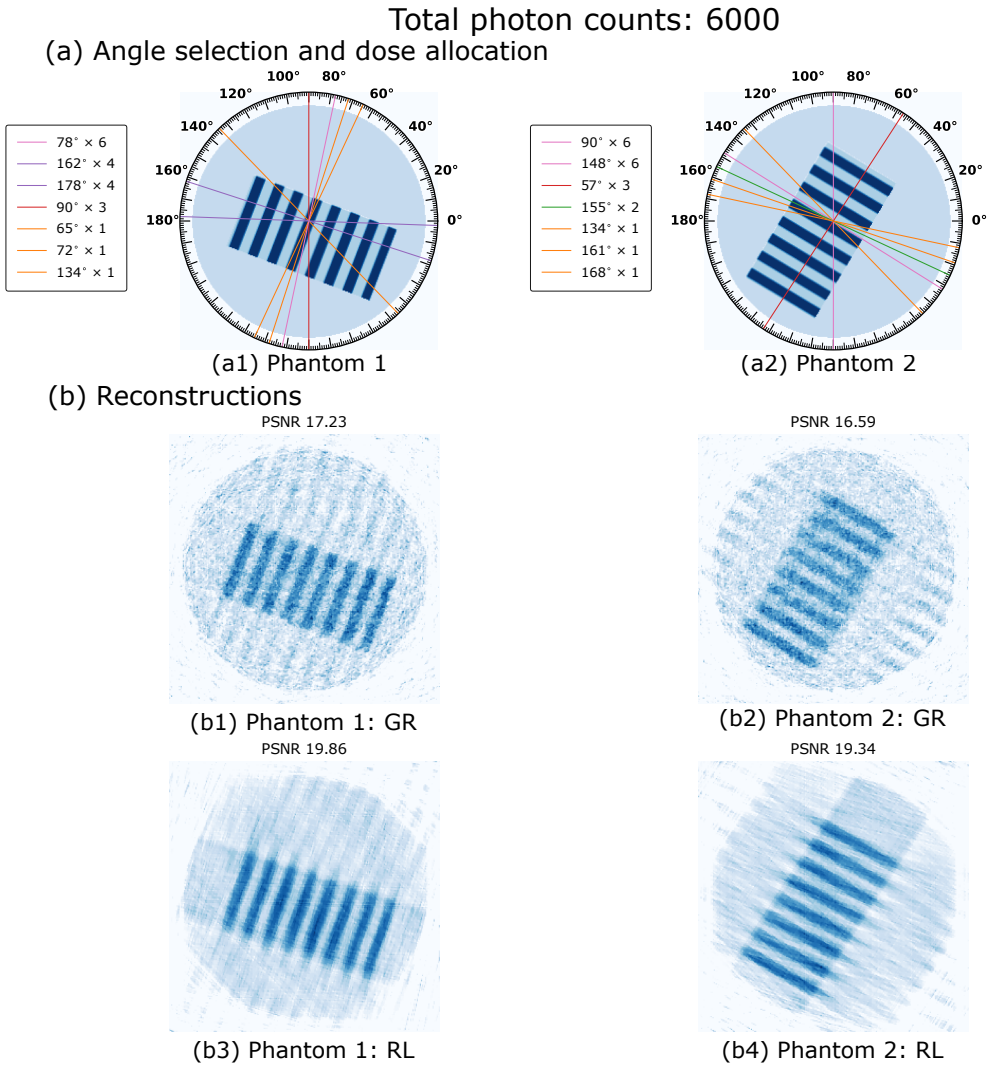


Figure 5.5: (a) Angle selection and dose allocation at a total dose of 6000 photons for wedge-shaped phantoms (a1–a2). Each colored spoke represents a parallel-beam view at angle θ, measured counter-clockwise from the x-axis; angles θ and $\theta + 180^\circ$ are equivalent. Colors indicate the number of repeated selections and thus the allocated photons at that angle. With 6000 photons in total, each selection corresponds to 300 photons; for example, an orange spoke denotes one selection (300 photons), and a green spoke denotes two selections (600 photons). (b) Reconstructions obtained under the GR (b1–b2) and RL (b3–b4) policies at the same total dose. The RL policy adaptively concentrates measurements in more informative directions, yielding improved reconstruction quality.

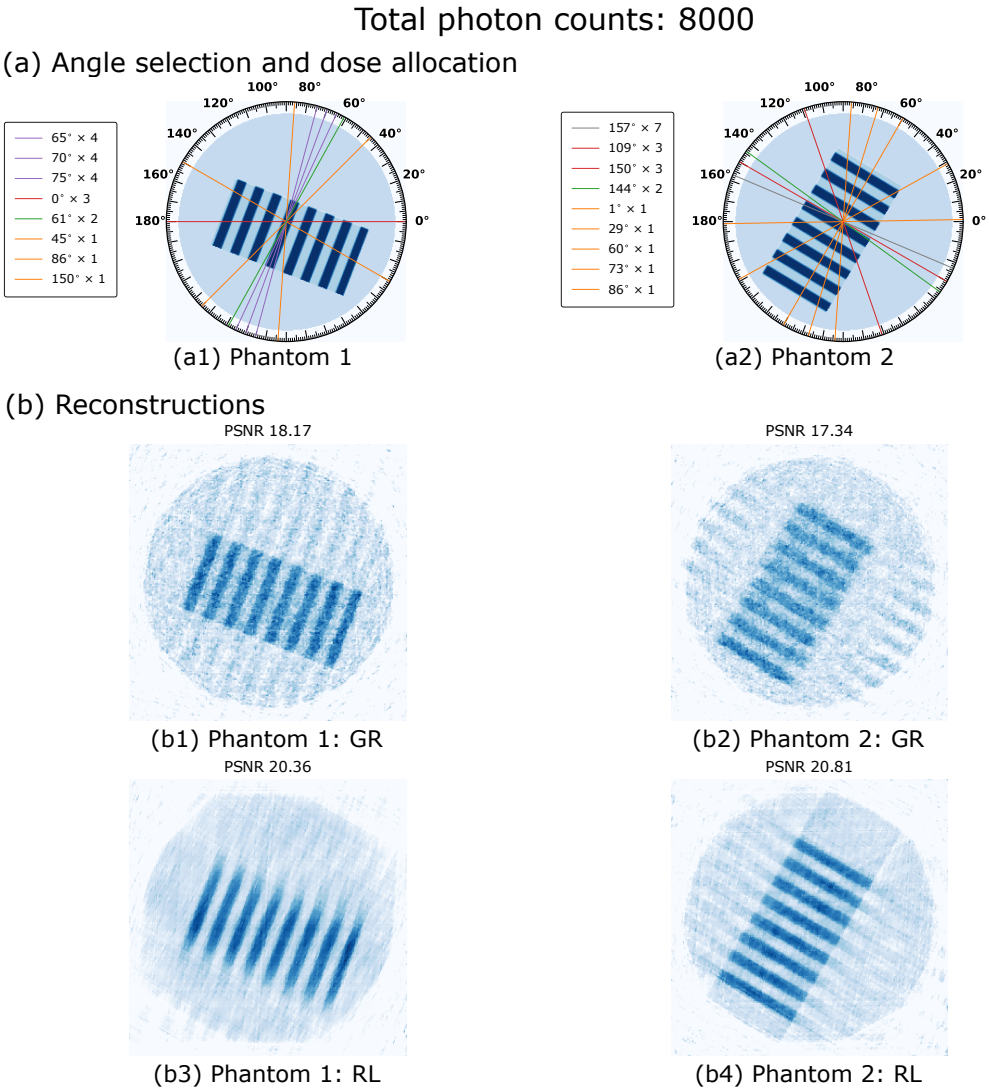


Figure 5.6: (a) Angle selection and dose allocation at a total dose of 8000 photons for wedge-shaped phantoms (a1–a2). Each colored spoke represents a parallel-beam view at angle θ, measured counter-clockwise from the x-axis; angles θ and $\theta + 180°$ are equivalent. Colors indicate the number of repeated selections and thus the allocated photons at that angle. With 8000 photons in total, each selection corresponds to 400 photons; for example, an orange spoke denotes one selection (400 photons), and a green spoke denotes two selections (800 photons). (b) Reconstructions obtained under the GR (b1–b2) and RL (b3–b4) policies at the same total dose. The RL policy adaptively concentrates measurements in more informative directions, yielding improved reconstruction quality.

select angles and allocate dose in a way that maintains reconstruction quality for non-defective phantoms while improving defect detectability. We still used the dose-aware PnP–PWLS with a TV prior as the reconstruction algorithm. The elliptical region is assumed to be defect-prone, and the ground-truth defect mask is used during training. Accordingly, we define an ROI covering the ellipse and compute rewards within this region. Figure (5.7) illustrates the ROI, the ground-truth defect mask, and the surrounding background region used for computing the CNR in Equation (5.18).

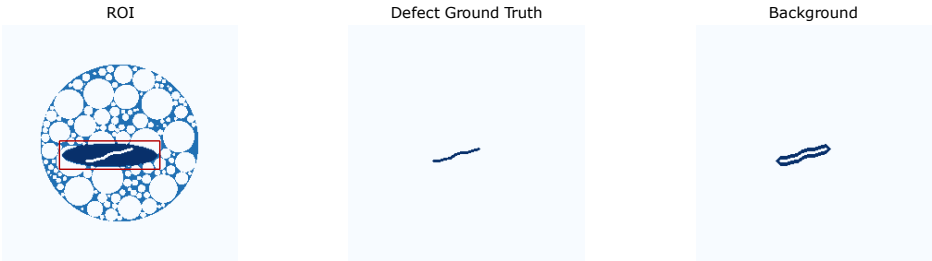


Figure 5.7: Illustration of the regions used for defect detection. (Left) ROI highlighted by a red rectangle. (Middle) Ground-truth defect mask. (Right) Background region used for CNR calculation.

Training procedure. The agent was trained on a dataset of 300 phantoms, comprising 150 non-defective and 150 defective cases. Three separate RL policies were trained for three different total photon budgets of 1000, 1500, and 2000. Training was performed for 2000 episodes using the reward function defined in Eq. (5.22). The weight parameters α and β in Eq. (5.22) were set to 0.2 and 3.5, respectively, to balance the two tasks.

$$R(\bar{\boldsymbol{x}}, \widehat{\boldsymbol{x}}) = \begin{cases} \alpha \, \mathrm{PSNR}_{\Omega}(\bar{\boldsymbol{x}}, \widehat{\boldsymbol{x}}), & \text{if } \bar{\boldsymbol{x}} \text{ is non-defective,} \\ \beta \, \mathrm{CNR}_{\mathcal{D},\mathcal{B}}(\bar{\boldsymbol{x}}, \widehat{\boldsymbol{x}}), & \text{if } \bar{\boldsymbol{x}} \text{ is defective,} \end{cases} \tag{5.22}$$

where α and β control the relative importance of reconstruction accuracy and defect detectability.

Validation procedure. Evaluation was carried out on a separate test set of 200 phantoms with unseen rotation angles, including 100 non-defective and 100 defective cases. All results reported in this section are obtained from this validation set using the trained policy.

Quantitative comparison. Table 5.2 summarizes the performance of the GR, PU, and RL policies on the unseen-rotation test set across total photon budgets of 1000, 1500, and 2000. For non-defective phantoms, the RL policy achieves PSNR comparable to or higher than GR at 1000 and 1500 photons, with significant mean improvements of 0.56 and 0.06 dB, respectively, based on paired bootstrap confidence intervals and Benjamini–Hochberg FDR correction. At 2000 photons, the RL policy obtains a slightly lower PSNR than GR, with a mean difference of -0.26 dB. For defective phantoms, however, the RL policy consistently achieves substantially higher CNR than GR at all dose levels. The mean CNR improvements are 0.47, 0.61, and 0.59 for total photon budgets of 1000, 1500, and 2000, respectively, and all corresponding 95% paired bootstrap confidence intervals are strictly positive after FDR correction. The PU baseline does not outperform GR in either PSNR or CNR, indicating that simple myopic uncertainty-based adaptivity is insufficient for this defect-detection task. Overall, these results show that the learned RL policy prioritizes task-specific defect detectability while largely preserving reconstruction quality.

Table 5.2: Performance comparison of GR, projected-uncertainty (PU), and RL policies on unseen-rotation tests for the foam-defect experiment. Results are reported as mean ± standard deviation. The RL–GR difference is reported with 95% paired bootstrap confidence intervals. One-sided p-values for RL outperforming GR were corrected using Benjamini–Hochberg FDR control at $\alpha = 0.05$.

Total photons	Policy	PSNR (Non-def.)	CNR (Def.)	Average number of angles	PSNR diff. RL–GR [95% CI]	CNR diff. RL–GR [95% CI]
1000	GR	19.50 ± 1.28	0.51 ± 0.25	10		
	PU	19.19 ± 1.30	0.49 ± 0.39	10	0.56 [0.52, 0.60]*	0.47 [0.45, 0.50]*
	RL	20.06 ± 1.52	0.98 ± 0.42	6.92 ± 1.41		
1500	GR	20.68 ± 1.23	0.62 ± 0.30	10		
	PU	19.79 ± 1.35	0.59 ± 0.46	10	0.06 [0.02, 0.10]*	0.61 [0.59, 0.63]*
	RL	20.74 ± 1.49	1.23 ± 0.42	7.15 ± 1.27		
2000	GR	21.38 ± 1.20	0.70 ± 0.32	10		
	PU	20.10 ± 1.44	0.62 ± 0.49	10	-0.26 [-0.30, -0.22]	0.59 [0.57, 0.62]*
	RL	21.12 ± 1.50	1.29 ± 0.47	7.56 ± 1.25		

*Significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$.

Angle selection and reconstruction analysis. Figures (5.8) and (5.9) (a1-a4) illustrate the angle selections produced by the GR and RL policies for non-defective and defective phantoms at total doses of 1000 and 2000 photons, respectively. While the GR policy distributes views uniformly regardless of phantom type, the RL policy adapts its strategy: for defective phantoms, angles and dose are concentrated around the defect region. Comparing across dose levels shows that higher photon budgets

lead to broader angular coverage, while the presence of defects drives the RL policy to focus sampling more locally.

The ultimate goal of this experiment is to evaluate whether adaptive angle and dose selection can improve defect detectability. Figures (5.8) and (5.9) (b1-b4) show reconstruction results for defective phantoms at total doses of 1000 and 2000 photons, respectively. Reconstructions using the GR policy fail to make the defects clearly visible. In contrast, reconstructions obtained with the RL policy consistently achieve higher CNR values and reveal defects more distinctly, validating the benefit of dose-aware adaptive view selection for defect detection.

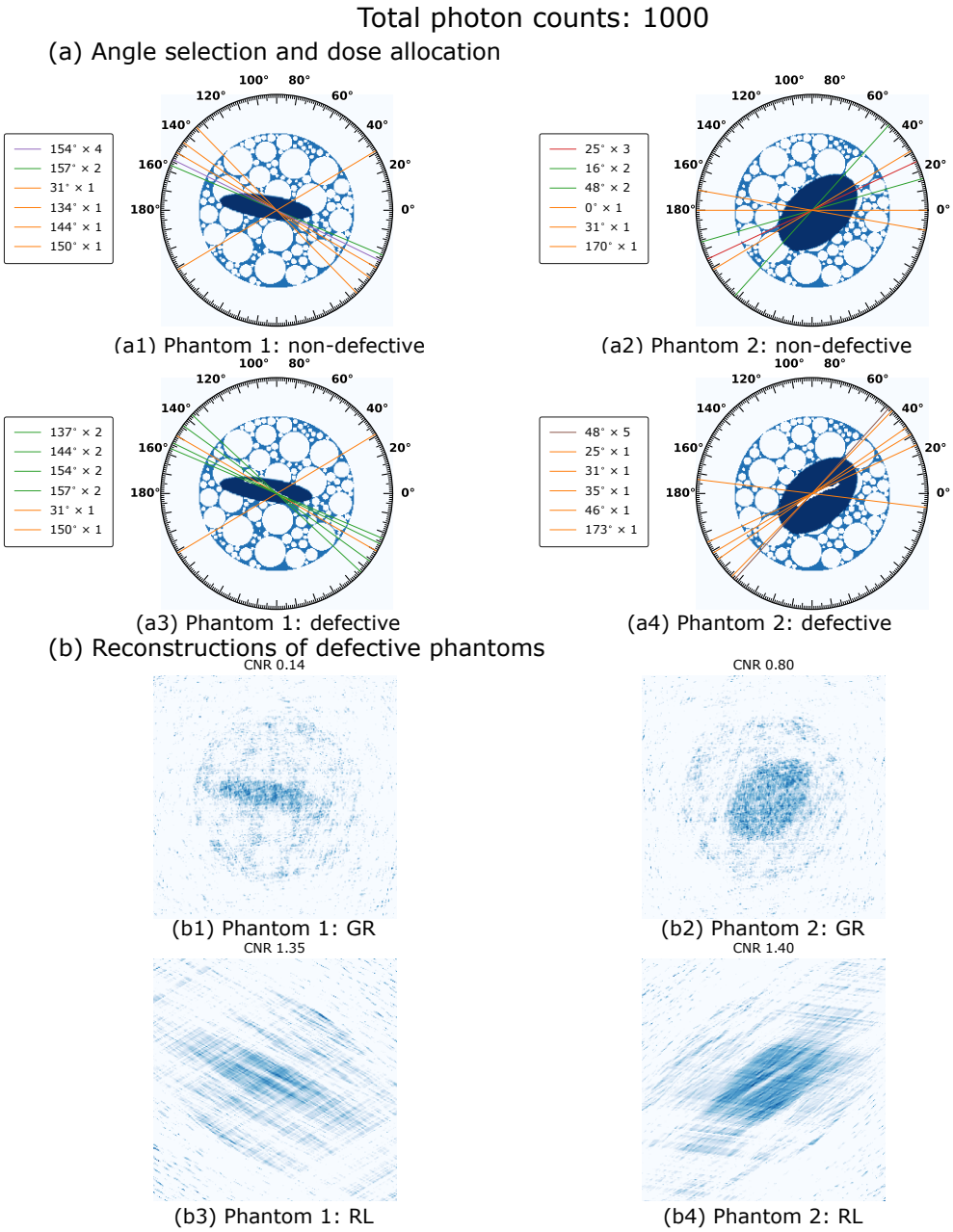


Figure 5.8: (a) Angle selection and dose allocation at a total dose of 1000 photons for non-defective (a1–a2) and defective (a3–a4) phantoms. Colored spokes denote projection angles, with color indicating the number of repeats (i.e., relative photon allocation). An orange spoke corresponds to one selection (base level: 100 photons), while colors such as green indicate multiple repeats and therefore higher allocated photons. (b) Reconstructions of defective phantoms using the GR (b1–b2) and RL (b3–b4) policies. The RL policy adaptively concentrates measurements near the defect, yielding higher CNR and improved defect visibility compared with the fixed GR policy.

Computational cost. To quantify the training cost without repeating the full training procedure, we benchmarked the runtime of representative training runs and measured the average wall-clock time per training episode. The benchmark was performed on a single NVIDIA L40 GPU with 44.39 GB memory, using PyTorch 2.11.0+cu130 and CUDA 13.0. Since each training episode contains the same acquisition length and reconstruction procedure as in the full training, the total training cost was computed by multiplying the measured average per-episode runtime by the number of training episodes. For the wedge phantom experiment, 50 representative training episodes required 4.28 hours, corresponding to an average of 0.0856 hours per episode. Multiplying this by the full training schedule of 1500 episodes gives a training cost of approximately 128.33 GPU-hours. For the foam-defect experiment, 50 representative training episodes required 4.40 hours, corresponding to an average of 0.0880 hours per episode. Multiplying this by the full training schedule of 2000 episodes gives a training cost of approximately 176.06 GPU-hours. Because each run used a single GPU, the reported GPU-hours are equal to the wall-clock hours.

Summary. These experiments demonstrate that the proposed RL policy adaptively selects relative informative angles and allocates dose across projection angles according to phantom content. It preserves reconstruction quality for non-defective phantoms while substantially improving defect detectability in defective cases, consistently outperforming the GR policy in both CNR and visual clarity.

5.5 Discussion

These results demonstrate that the proposed dose-aware reconstruction algorithms are particularly effective when dose is distributed unequally across projection angles. Moreover, the RL policy achieves superior reconstruction quality and defect detectability by leveraging fewer angles and adaptively distributing dose within a fixed budget, compared to the GR policy, which relies on more angles with equal dose allocation.

The dose-aware PWLS formulation is based on a standard second-order approximation of the Poisson negative log-likelihood; the main novelty of this work lies in integrating this reconstruction model into an adaptive RL framework for joint angle selection and dose allocation. The angle-dependent weights provide a statistically motivated way to account for dose-dependent noise by assigning higher confidence to high-dose projections and lower confidence to noisier low-dose projections. For a fixed acquisition sequence, the PnP–PWLS update behaves as a regularized weighted least-

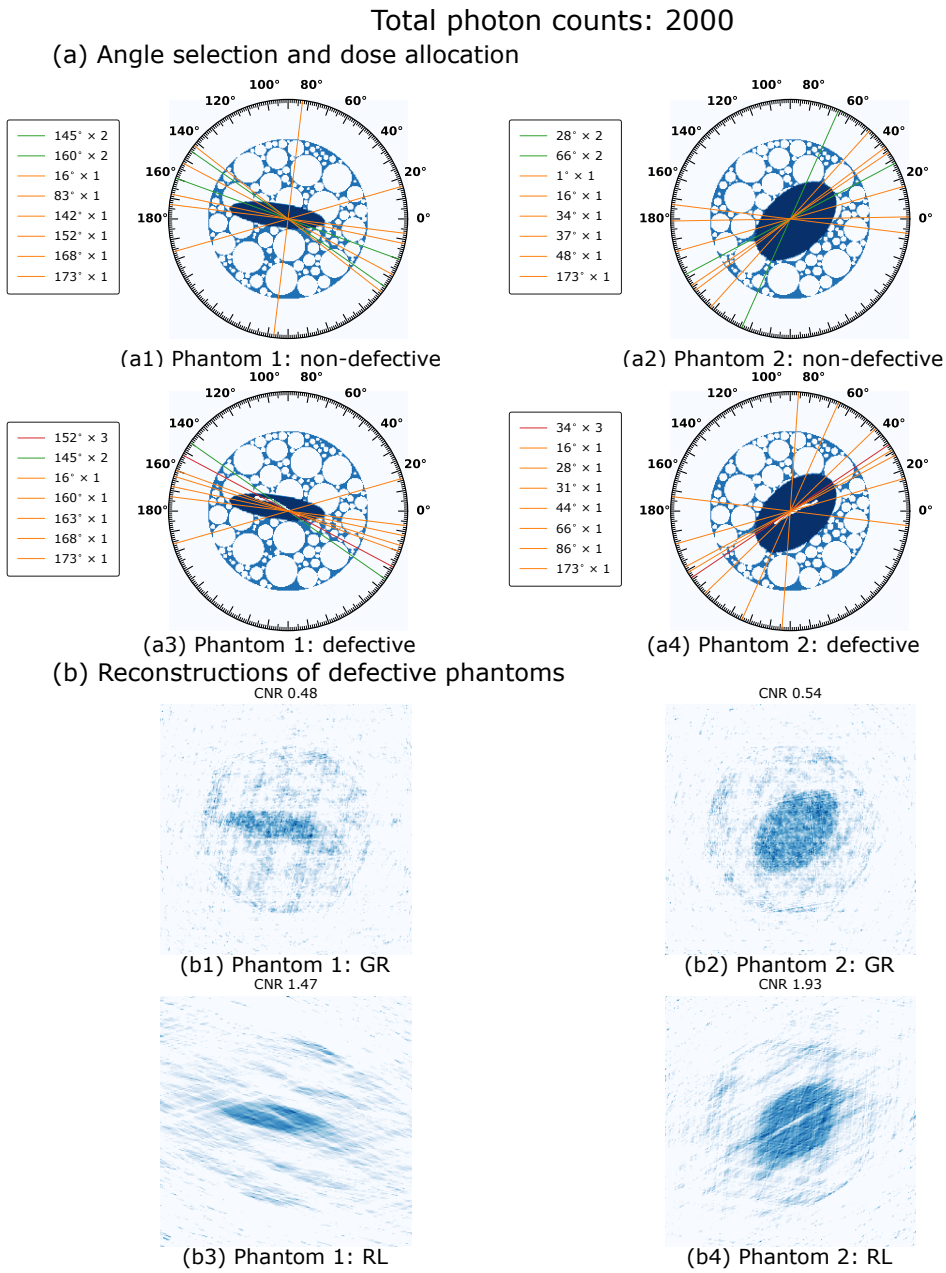


Figure 5.9: (a) Angle selection and dose allocation at a total dose of 2000 photons for non-defective (a1–a2) and defective (a3–a4) phantoms. Colored spokes denote projection angles, with color indicating the number of repeats (i.e., relative photon allocation). An orange spoke corresponds to one selection (base level: 200 photons), while colors such as green indicate multiple repeats and therefore higher allocated photons. (b) Reconstructions of defective phantoms using the GR (b1–b2) and RL (b3–b4) policies. The RL policy adaptively concentrates measurements near the defect, yielding higher CNR and improved defect visibility compared with the fixed GR policy.

squares iteration, and the estimated Lipschitz constant is used to stabilize the gradient step. However, because the RL policy changes the selected angles and dose distribution sequentially, the complete adaptive acquisition–reconstruction loop does not optimize a single fixed objective, and we therefore do not claim a formal global convergence guarantee. The reward design also influences the learned acquisition strategy. PSNR- and CNR-based rewards are used only during offline training, where ground-truth phantoms are available; the trained policy does not require ground truth during inference. The weights α and β control the trade-off between reconstruction quality and defect detectability, making the learned policy task-dependent. Consequently, the policy is expected to generalize within the object class and imaging conditions represented in the training distribution, but different tasks, datasets, or imaging conditions may require retraining, fine-tuning, or an expanded training distribution. Alternative reward formulations based on data consistency, uncertainty reduction, expected information gain, or downstream task confidence may further reduce reliance on ground-truth-based rewards and are important directions for future work.

Despite these encouraging findings, several limitations remain. First, only a Poisson noise model was considered, whereas real CT measurements are often better described by a mixture of Gaussian and Poisson noise [2, 36]. Accurate modeling of the noise characteristics of a specific CT system will be important for translating this approach to real data. Second, dose allocation is represented through repeated selection of the same projection angle. This provides a convenient abstraction for concentrating photons under a fixed dose budget, but practical implementation may require scanner-specific mechanisms such as exposure-time control, tube-current modulation, or repeated measurements. Third, extending the framework to three-dimensional datasets introduces additional challenges: iterative reconstruction becomes more computationally demanding, and the state and action spaces for RL grow substantially. Fourth, the validation experiments use unseen rotations, shifts, and shape variations, indicating generalization within the training distribution. However, we do not claim generalization to arbitrary out-of-distribution geometries or unrelated object classes. In practical industrial CT applications, this assumption is often reasonable because prior information about the inspected object class is available through CAD models, blueprints, design specifications, or previous scans. For substantially different geometries or imaging conditions, retraining, fine-tuning, or domain-randomized training would be required. Finally, future work may explore the integration of downstream tasks such as segmentation and classification, which could further highlight the practical benefits of adaptive angle and dose allocation.

5.6 Conclusion

In this paper, we proposed dose-aware reconstruction methods and developed a reinforcement learning strategy for joint angle selection and dose allocation. Our results demonstrate that explicitly modeling angle-dependent dose variation improves reconstruction quality and defect detectability under unequal dose distributions. Furthermore, the proposed RL-based framework shows the potential to reduce the number of projection angles while maintaining or improving image quality, thereby enabling more efficient and task-adaptive CT scanning.

Data availability statement

The data that support the findings of this study are available from the corresponding author, T.W., upon reasonable request. The implementation and synthetic datasets used in this study are publicly available on GitHub: `https://github.com/tianyuan1wang/ct-angle-dose-rl` and Zenodo: `https://doi.org/10.5281/zenodo.20486336`.

Chapter 6

Conclusion and Outlook

6.1 Contributions and Limitations

The primary objective of this thesis was to develop methods for adaptive projection angle selection in X-ray CT, with a particular focus on inline industrial inspection. This problem can naturally be formulated within a sequential OED framework, in which measurement decisions are updated based on previously acquired data.

However, directly applying classical sequential OED formulations to adaptive CT scanning poses several fundamental challenges. First, the sequential OED framework typically requires evaluating expectations over both the prior distribution and the data likelihood, which are generally unavailable in closed form for realistic CT systems. Second, practical CT problems are inherently high-dimensional, since both the unknown object and the measurement data are large-scale. Although approximate models for these probability distributions can be constructed, the resulting computations are often prohibitively expensive, limiting the applicability of classical sequential OED methods in time-critical industrial settings.

6.1.1 Contributions and Limitations

Chapter 2 formulates sequential OED for X-ray CT as a Partially Observable Markov Decision Process (POMDP) and addresses it using an Actor–Critic RL framework. A model-free formulation removes the requirement for closed-form expressions of the prior distribution and data likelihood, enables offline training, and allows the trained policy to be deployed during scanning. A series of two-dimensional phantoms, ranging

from simple to moderately complex shapes, were designed to validate the approach. For simple shapes such as triangles and pentagons, the RL agent learned to select projection angles tangential to object boundaries, which are known to be highly informative. Notably, a model trained on a mixture of different shapes retained the ability to select such informative angles. For a more complex modified Shepp–Logan phantom with rotatable ellipses, the agent successfully tracked these rotations by adapting its angle selection. The generalisability of the trained policy was also examined, showing robustness when transferring from clean to noisier data, and limited generalisation across substantially different shapes and intensity distributions.

Despite demonstrating that high-quality reconstructions can be obtained from very few projections in simple phantoms, the experimental setup remains idealised relative to industrial practice. In real industrial inspection, the primary interest often lies in the interpretability or task relevance of the reconstruction rather than purely numerical error metrics. Moreover, the parallel-beam geometry considered in this chapter is rarely used in industrial CT systems.

Building on the previous chapter, Chapter 3 extends the framework to task-specific angle selection rather than focusing solely on general reconstruction quality. Two types of defects—cracks and pores—were inserted into a predefined Region-of-Interest (RoI) of a simulated industrial component, representing areas prone to failure. The reward function was redesigned to incorporate defect detectability measures, including contrast-based metrics and segmentation performance quantified by the Dice score. As a result, the RL agent adaptively selected angles that improved defect visibility while maintaining overall reconstruction quality.

Although promising results were obtained for simulated industrial parts, it remains unclear how well the approach would transfer to real measurement data. In practice, noise and reconstruction artifacts may obscure small defects and reduce detectability. Furthermore, segmentation in this chapter relied on K-means clustering and assumed a single-material object, which is a strong simplification and not representative of many real industrial components.

Chapter 4 investigates optimal stopping within the sequential OED framework using an Actor–Critic approach. A terminal policy was introduced to determine when data acquisition should stop, allowing the number of projection angles to be selected adaptively. This terminal policy was optimized jointly with the angle-selection policy to maximize the expected cumulative reward. In addition, an experimental dataset was collected at the FleX-ray laboratory of Centrum Wiskunde & Informatica (CWI) by varying the tube current to generate different noise levels. The trained policy was

evaluated on experimental data to assess whether a model learned from simulations could transfer to real measurements. The results show that joint optimization of angle selection and stopping outperforms naive fixed stopping criteria, and that the trained model retains functionality on experimental data with similar characteristics. The terminal policy also adapts the number of projections to different object shapes and noise conditions.

Nevertheless, the experimental objects considered in this chapter were geometrically simple. Furthermore, the simulation model could be made more realistic. Only Gaussian noise was considered in the simulations, whereas the experimental data exhibit a combination of Gaussian and Poisson noise.

Chapter 5 extends the adaptive design framework by introducing dose level as an additional design parameter, directly influencing Poisson noise statistics. The RL agent was trained to jointly select projection angles and allocate dose levels adaptively. Task-specific acquisition was also investigated, with the agent allocating more projections and higher dose to directions aligned with crack-like defects. To enable reconstruction under non-uniform dose allocation, the conventional Penalized Weighted Least Squares (PWLS) formulation was modified to incorporate angle-dependent weighting. This dose-aware PWLS term was combined with a TV prior in a PnP framework, providing robust reconstructions under highly heterogeneous dose distributions.

In this chapter, dose allocation was implemented through repeated selection of projection angles, with higher repetition corresponding to higher dose. In principle, the framework could be extended to output angles and dose levels jointly and continuously after sufficient information has been gathered. Unlike Chapter 4, only Poisson noise was considered in this study, whereas real CT measurements typically involve a mixture of Poisson and Gaussian noise.

6.1.2 Outlook

Future work should extend the methods developed in this thesis from two-dimensional to fully three-dimensional geometries, as cone-beam CT is commonly used in industrial inspection. In addition, acquisition trajectories could be made more flexible by allowing zoom-in and zoom-out motions, rather than restricting the scan to planar trajectories. Achieving such flexible trajectories naturally requires the use of robotic manipulators to control the CT scanner or the object, and it would therefore be valuable to more tightly integrate robotic control with trajectory generation.

Since industrial inspection ultimately aims to extract actionable information from CT reconstructions, downstream image analysis tasks remain an important direction for further research. For example, in Chapter 3, segmentation was performed using relatively simple methods. More advanced approaches, such as deep learning-based segmentation, could be incorporated, allowing angle selection strategies to be optimized directly for improved segmentation or defect detection performance. Similarly, the task-specific optimal stopping strategy studied in Chapter 4 could be extended to task-driven stopping criteria, where scanning terminates once a defect has been detected with sufficient confidence.

From a theoretical perspective, future work could aim to make the current *black-box* formulation more transparent by incorporating more accurate physical models. This includes developing improved forward and inverse models, as well as more realistic noise models. As discussed in Chapters 4 and 5, CT measurements are often affected by a combination of Poisson and Gaussian noise, and modeling these effects more accurately—potentially in a system-specific manner—could improve both reconstruction quality and policy training. More realistic simulation models would also facilitate better transfer of policies trained on simulated data to experimental or industrial data.

By combining these practical and theoretical developments, a physics-based simulation environment could be constructed that integrates X-ray imaging and robotic motion in a manner closely resembling real-world inspection systems. Such an environment would reduce the need for extensive data collection in laboratory settings by enabling the generation of realistic synthetic data. Within this framework, robotic systems could learn efficient and informative acquisition strategies, ultimately supporting fully adaptive and efficient inline industrial inspection.

Bibliography

[1] Agin. "Computer vision systems for industrial inspection and assembly". In: *Computer* 13.5 (1980), pp. 11–20.

[2] Vladyslav Andriiashen, Robert van Liere, Tristan van Leeuwen, and K. Joost Batenburg. "X-ray Image Generation as a Method of Performance Prediction for Real-Time Inspection: A Case Study". In: *Journal of Nondestructive Evaluation* 43.3 (2024), p. 79.

[3] Vladyslav Andriiashen, Robert van Liere, Tristan van Leeuwen, and Kees Joost Batenburg. "Unsupervised foreign object detection based on dual-energy absorptiometry in the food industry". In: *Journal of Imaging* 7.7 (2021), p. 104.

[4] Vladyslav Viacheslavovych Andriiashen. "Data-driven X-ray image generation for industrial applications". English. Proefschrift. PhD thesis. Eindhoven University of Technology, May 2025. ISBN: 978-90-386-6364-7.

[5] Pierre-Luc Bacon, Jean Harb, and Doina Precup. "The option-critic architecture". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[6] Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. "Computed tomography reconstruction using deep image prior and learned reconstruction methods". In: *Inverse Problems* 36.9 (2020), p. 094004.

[7] Riccardo Barbano, Johannes Leuschner, Javier Antorán, Bangti Jin, and José Miguel Hernández-Lobato. "Bayesian experimental design for computed tomography with the linearised deep image prior". In: *arXiv preprint arXiv:2207.05714* (2022).

[8] K Joost Batenburg, Wagner Fortes, Lajos Hajdu, and Robert Tijdeman. "Bounds on the difference between reconstructions in binary tomography". In: *International Conference on Discrete Geometry for Computer Imagery*. Springer. 2011, pp. 369–380.

[9] K Joost Batenburg, Willem Jan Palenstijn, Péter Balázs, and Jan Sijbers. "Dynamic angle selection in binary tomography". In: *Computer Vision and Image Understanding* 117.4 (2013), pp. 306–318.

[10] Fabian Bauer, Daniel Forndran, Thomas Schromm, and Christian U Grosse. "Practical part-specific trajectory optimization for robot-guided inspection via computed tomography". In: *Journal of Nondestructive Evaluation* 41.3 (2022), p. 55.

[11] Sebastian Becker, Patrick Cheridito, and Arnulf Jentzen. "Deep optimal stopping". In: *Journal of Machine Learning Research* 20.74 (2019), pp. 1–25.

[12] Simon Bellens, Patricio Guerrero, Patrick Vandewalle, and Wim Dewulf. "Machine learning in industrial X-ray computed tomography–a review". In: *CIRP Journal of Manufacturing Science and Technology* 51 (2024), pp. 324–341.

[13] Tom Blau, Edwin V Bonilla, Iadine Chades, and Amir Dezfouli. "Optimizing sequential experimental design with deep reinforcement learning". In: *International conference on machine learning*. PMLR. 2022, pp. 2107–2128.

[14] Martin Burger, Andreas Hauptmann, Tapio Helin, Nuutti Hyvönen, and Juha-Pekka Puska. "Sequentially optimized projections in X-ray imaging". In: *Inverse Problems* 37.7 (2021), p. 075006.

[15] Victor Bussy, Caroline Vienne, Julie Escoda, and Valérie Kaftandjian. "Best projections selection algorithm based on constrained QDEIM for sparse-views X-ray Computed Tomography". In: (2022).

[16] Victor Bussy, Caroline Vienne, and Valérie Kaftandjian. "Fast algorithms based on Empirical Interpolation Methods for selecting best projections in Sparse-View X-ray Computed Tomography using a priori information". In: *NDT & E International* 134 (2023), p. 102768.

[17] Thorsten M Buzug. *Computed tomography*. Springer, 2011.

[18] Simone Carmignato, Wim Dewulf, Richard Leach, et al. *Industrial X-ray computed tomography*. Vol. 10. Springer, 2018.

[19] Kathryn Chaloner and Isabella Verdinelli. "Bayesian experimental design: A review". In: *Statistical science* (1995), pp. 273–304.

[20] Antonin Chambolle. “An algorithm for total variation minimization and applications”. In: *Journal of Mathematical imaging and vision* 20.1 (2004), pp. 89–97.

[21] Timothy M Craig, Ajinkya A Kadu, Kees Joost Batenburg, and Sara Bals. “Real-time tilt undersampling optimization during electron tomography of beam sensitive samples using golden ratio scanning and RECAST3D”. In: *Nanoscale* 15.11 (2023), pp. 5391–5402.

[22] Andrei Dabravolski, Kees Joost Batenburg, and Jan Sijbers. “Dynamic angle selection in X-ray computed tomography”. In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 324 (2014), pp. 17–24.

[23] Niranjan Damera Venkata and Chiranjib Bhattacharyya. “Deep recurrent optimal stopping”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 12222–12244.

[24] Gilles Demaurex and Laurent Sallé. “Detection of physical hazards”. In: *Food safety management*. Elsevier, 2014, pp. 511–533.

[25] Wim Dewulf, Harald Bosse, Simone Carmignato, and Richard Leach. “Advances in the metrological traceability and performance of X-ray computed tomography”. In: *CIRP annals* 71.2 (2022), pp. 693–716.

[26] Noam Elata, Tomer Michaeli, and Michael Elad. “Adaptive compressed sensing with diffusion-based posterior sampling”. In: *European Conference on Computer Vision*. Springer. 2025, pp. 290–308.

[27] Tommy Elfving, Per Christian Hansen, and Touraj Nikazad. “Semiconvergence and relaxation parameters for projected SIRT algorithms”. In: *SIAM Journal on Scientific Computing* 34.4 (2012), A2000–A2017.

[28] Valerii Fedorov. “Optimal experimental design”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.5 (2010), pp. 581–589.

[29] Andreas Fischer, Tobias Lasser, Michael Schrapp, Jürgen Stephan, and Peter B Noël. “Object specific trajectory optimization for industrial X-ray computed tomography”. In: *Scientific reports* 6.1 (2016), pp. 1–9.

[30] Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. “Deep adaptive design: Amortizing sequential bayesian experimental design”. In: *International conference on machine learning*. PMLR. 2021, pp. 3384–3395.

[31] Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. "Variational Bayesian optimal experimental design". In: *Advances in neural information processing systems* 32 (2019).

[32] Adam Evan Foster. "Variational, Monte Carlo and policy-based approaches to Bayesian experimental design". PhD thesis. University of Oxford, 2021.

[33] Fraunhofer Institute for Integrated Circuits IIS. *Inline CT: Computed Tomography within Production*. URL: `https://www.iis.fraunhofer.de/en/ff/zfp/tech/inline-computertomographie.html` (visited on 07/29/2026).

[34] Peter I Frazier. "A tutorial on Bayesian optimization". In: *arXiv preprint arXiv:1807.02811* (2018).

[35] Peter I Frazier. "Bayesian optimization". In: *Recent advances in optimization and modeling of contemporary problems*. Informs, 2018, pp. 255–278.

[36] Adriaan Graas and Felix Lucka. "Scintillator decorrelation for self-supervised x-ray radiograph denoising". In: *Measurement Science and Technology* 36.6 (2025), p. 065415.

[37] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning". In: *arXiv preprint arXiv:2501.12948* (2025).

[38] Per Christian Hansen, Jakob Jørgensen, and William RB Lionheart. *Computed Tomography: Algorithms, Insight, and Just Enough Theory*. SIAM, 2021.

[39] Tapio Helin, Nuutti Hyvönen, and Juha-Pekka Puska. "Edge-promoting adaptive Bayesian experimental design for X-ray imaging". In: *SIAM Journal on Scientific Computing* 44.3 (2022), B506–B530.

[40] Allard A Hendriksen, Dirk Schut, Willem Jan Palenstijn, Nicola Viganó, Jisoo Kim, Daniël M Pelt, Tristan Van Leeuwen, and Kees Joost Batenburg. "Tomosipo: fast, flexible, and convenient 3D tomography for complex scanning geometries in Python". In: *Optics Express* 29.24 (2021), pp. 40494–40513.

[41] Allard Adriaan Hendriksen, Daniël Maria Pelt, and K Joost Batenburg. "Noise2inverse: Self-supervised deep convolutional denoising for tomography". In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 1320–1335.

[42] Gabriel Herl, Jochen Hiller, Mareike Thies, Jan-Nico Zaech, Mathias Unberath, and Andreas Maier. "Task-specific trajectory optimisation for twin-robotic X-ray tomography". In: *IEEE Transactions on Computational Imaging* 7 (2021), pp. 894–907.

[43] Xun Huan, Jayanth Jagalur, and Youssef Marzouk. "Optimal experimental design: Formulations and computations". In: *Acta Numerica* 33 (2024), pp. 715–840.

[44] Shengyi Huang and Santiago Ontañón. "A closer look at invalid action masking in policy gradient algorithms". In: *arXiv preprint arXiv:2006.14171* (2020).

[45] Daniil Kazantsev, Folkert Bleichrodt, Tristan van Leeuwen, Anders Kaestner, Philip J Withers, Kees Joost Batenburg, and Peter D Lee. "A novel tomographic reconstruction method based on the robust Student's t function for suppressing data outliers". In: *IEEE Transactions on Computational Imaging* 3.4 (2017), pp. 682–693.

[46] IG Kazantsev. "Information content of projections". In: *Inverse problems* 7.6 (1991), p. 887.

[47] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[48] Maximilian B Kiss, Sophia B Coban, K Joost Batenburg, Tristan van Leeuwen, and Felix Lucka. "2DeteCT-A large 2D expandable, trainable, experimental Computed Tomography dataset for machine learning". In: *Scientific data* 10.1 (2023), p. 576.

[49] Thomas Kohler. "A projection access scheme for iterative reconstruction based on the golden section". In: *IEEE Symposium Conference Record Nuclear Science 2004.* Vol. 6. IEEE. 2004, pp. 3961–3965.

[50] Andreas Krause and Daniel Golovin. "Submodular function maximization." In: *Tractability* 3.71-104 (2014), p. 3.

[51] Tristan van Leeuwen, Christoph Brune, and Marcello Carioni. "An invertible generative model for forward and inverse problems". In: *arXiv preprint arXiv:2509.03910* (2025).

[52] Kacper Libera, Dirk Schut, Effrosyni Kritsi, Louis van Steijn, Timothy Dallman, and Len Lipman. "Detecting wing fractures in chickens using deep learning, photographs and computed tomography scanning". In: *Poultry Science* (2025), p. 105264.

[53] Dennis Victor Lindley. *Bayesian statistics: A review*. SIAM, 1972.

[54] Annabel Matz, Wolfgang Holub, and Richard Schielein. "Trajectory optimization in computed tomography based on object geometry". In: *11th Conference on Industrial Computed Tomography, Wels, Austria (iCT 2022)*. 2022.

[55] Cynthia H McCollough, Michael R Bruesewitz, and James M Kofler Jr. "CT dose reduction and dose management tools: overview of available options". In: *Radiographics* 26.2 (2006), pp. 503–512.

[56] Toby J Mitchell. "An algorithm for the construction of "D-optimal" experimental designs". In: *Technometrics* 42.1 (2000), pp. 48–54.

[57] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[58] Jennifer L Mueller and Samuli Siltanen. *Linear and nonlinear inverse problems with practical applications*. SIAM, 2012.

[59] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. "An analysis of approximations for maximizing submodular set functions—I". In: *Mathematical programming* 14.1 (1978), pp. 265–294.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.

[61] Daniël M Pelt, Allard A Hendriksen, and Kees Joost Batenburg. "Foam-like phantoms for comparing tomography algorithms". In: *Synchrotron Radiation* 29.1 (2022), pp. 254–265.

[62] "Penalized weighted least-squares image reconstruction for positron emission tomography". In: *IEEE transactions on medical imaging* 13.2 (1994), pp. 290–300.

[63] Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.

[64] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. "Modern Bayesian experimental design". In: *Statistical Science* 39.1 (2024), pp. 100–114.

[65] Leonid I Rudin, Stanley Osher, and Emad Fatemi. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.

[66] Lars Ruthotto, Julianne Chung, and Matthias Chung. “Optimal experimental design for inverse problems with state constraints”. In: *SIAM Journal on Scientific Computing* 40.4 (2018), B1080–B1100.

[67] EG Ryan, CC Drovandi, JM McGree, and AN Pettitt. “Fully Bayesian optimal experimental design: A review”. In: *International Statistical Review* 84 (2016), pp. 128–154.

[68] Richard Schielein. “Analytische Simulation und Aufnahmeplanung für die industrielle Röntgencomputertomographie”. In: *Dissertation* (2018).

[69] Linda-Sophie Schneider, Mareike Thies, Richard Schielein, Christopher Syben, Mathias Unberath, and Andreas Maier. “Learning-based Trajectory Optimization for a Twin Robotic CT System”. In: (2022).

[70] Thomas Schreiber, Christoph Netsch, Sören Eschweiler, Tianyuan Wang, Thomas Storek, Marc Baranski, and Dirk Müller. “Application of data-driven methods for energy system modelling demonstrated on an adaptive cooling supply system”. In: *Energy* 230 (2021), p. 120894.

[71] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. “High-dimensional continuous control using generalized advantage estimation”. In: *arXiv preprint arXiv:1506.02438* (2015).

[72] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[73] Dirk Elias Schut, Kees Joost Batenburg, Robert Van Liere, and Tristan Van Leeuwen. “TOP-CT: trajectory with overlapping projections X-ray computed tomography”. In: *IEEE Transactions on Computational Imaging* 8 (2022), pp. 598–608.

[74] Dirk Elias Schut, Rachael Maree Wood, Anna Katharina Trull, Rob Schouten, Robert van Liere, Tristan van Leeuwen, and Kees Joost Batenburg. “Joint 2D to 3D image registration workflow for comparing multiple slice photographs and CT scans of apple fruit with internal disorders”. In: *Postharvest Biology and Technology* 211 (2024), p. 112814.

[75] Matthias Seeger. “On the submodularity of linear experimental design”. In: *Unpublished Note* (2009).

[76] Kun Shao, Zhentao Tang, Yuanheng Zhu, Nannan Li, and Dongbin Zhao. “A survey of deep reinforcement learning in video games”. In: *arXiv preprint arXiv:1912.10944* (2019).

[77] Wanggang Shen and Xun Huan. “Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning”. In: *Computer Methods in Applied Mechanics and Engineering* 416 (2023), p. 116304.

[78] Ziju Shen, Yufei Wang, Dufan Wu, Xu Yang, and Bin Dong. “Learning to scan: A deep reinforcement learning approach for personalized scanning in CT imaging”. In: *Inverse Problems & Imaging* 16.1 (2022).

[79] Jiayang Shi, Daniël M Pelt, and K Joost Batenburg. “Multi-stage deep learning artifact reduction for parallel-beam computed tomography”. In: *Synchrotron Radiation* 32.2 (2025).

[80] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

[81] Mathias Staudigl, Simon Weissmann, and Tristan Van Leeuwen. “Derivative-free stochastic bilevel optimization for inverse problems”. In: *Computational Optimization and Applications* (2025), pp. 1–54.

[82] J Webster Stayman and Jeffrey H Siewerdsen. “Task-based trajectories in iteratively reconstructed interventional cone-beam CT”. In: *Proc. 12th Int. Meet. Fully Three-Dimensional Image Reconstr. Radiol. Nucl. Med* (2013), pp. 257–260.

[83] W Sun, SB Brown, and RK Leach. “An overview of industrial X-ray computed tomography.” In: (2012).

[84] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[85] Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. “Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction”. In: *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 2011, pp. 761–768.

[86] Ruben Rodriguez Torrado, Philip Bontrager, Julian Togelius, Jialin Liu, and Diego Perez-Liebana. "Deep reinforcement learning for general video game ai". In: *2018 IEEE conference on computational intelligence and games (CIG)*. IEEE. 2018, pp. 1–8.

[87] J Ultrasonics. "Industrial applications of ultrasound-a review ii. measurements, tests, and process control using low-intensity ultrasound". In: *IEEE Transactions on sonics and Ultrasonics* (1975).

[88] Enrico Ursella. "In-line industrial computed tomography applications and developments". In: (2021).

[89] Wim Van Aarle, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K Joost Batenburg, and Jan Sijbers. "Fast and flexible X-ray tomography using the ASTRA toolbox". In: *Optics express* 24.22 (2016), pp. 25129–25147.

[90] Wim Van Aarle, Willem Jan Palenstijn, Jan De Beenhouwer, Thomas Altantzis, Sara Bals, K Joost Batenburg, and Jan Sijbers. "The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography". In: *Ultramicroscopy* 157 (2015), pp. 35–47.

[91] László Varga, Péter Balázs, and Antal Nagy. "Projection selection dependency in binary tomography". In: *Acta Cybernetica* 20.1 (2011), pp. 167–187.

[92] Marcelo Venere, Hstau Liao, and Alejandro Clausse. "A genetic algorithm for adaptive tomography of elliptical objects". In: *IEEE Signal Processing Letters* 7.7 (2000), pp. 176–178.

[93] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. "Plug-and-play priors for model based reconstruction". In: *2013 IEEE global conference on signal and information processing*. IEEE. 2013, pp. 945–948.

[94] Bussy Victor, Caroline Vienne, and Valérie Kaftandjian. "Fast Algorithms Based on Empirical Interpolation Methods for Selecting Best Projections in Sparse-View X-Ray Computed Tomography Using a Priori Information". In: *Available at SSRN 4170496* ().

[95] Jing Wang, Tianfang Li, Hongbing Lu, and Zhengrong Liang. "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography". In: *IEEE transactions on medical imaging* 25.10 (2006), pp. 1272–1283.

[96] Tianyuan Wang, Virginia Florian, Richard Schielein, Christian Kretzer, Stefan Kasperl, Felix Lucka, and Tristan van Leeuwen. "Task-Adaptive Angle Selection for Computed Tomography-Based Defect Detection". In: *Journal of Imaging* 10.9 (2024), p. 208. DOI: 10.3390/jimaging10090208.

[97] Tianyuan Wang, Felix Lucka, and Tristan van Leeuwen. "Sequential Experimental Design for X-Ray CT Using Deep Reinforcement Learning". In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 953–968. DOI: 10.1109/TCI.2024.3414273.

[98] Tianyuan Wang, Felix Lucka, and Tristan van Leeuwen. *X-ray Computed Tomography Case Study: Triangle and Pentagon Datasets with Various Sizes, Scales, and Noise Levels*. Zenodo, Feb. 2025.

[99] Tianyuan Wang, Felix Lucka, Daniël M. Pelt, K. Joost Batenburg, and Tristan van Leeuwen. "Dynamic Angle Selection in X-Ray CT: A Reinforcement Learning Approach to Optimal Stopping". In: *Applied Mathematics for Modern Challenges* 5 (2025), pp. 36–63. DOI: 10.3934/ammc.2025010.

[100] Tianyuan Wang, Daniël M Pelt, Felix Lucka, Tristan van Leeuwen, and Kees Joost Batenburg. "Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction". In: *Applied Mathematics in Science and Engineering* 34.1 (2026), p. 2705107. DOI: 10.1080/27690911.2026.2705107.

[101] Wenjun Xia, Zexin Lu, Yongqiang Huang, Yan Liu, Hu Chen, Jiliu Zhou, and Yi Zhang. "CT reconstruction with PDF: Parameter-dependent framework for data from multiple geometries and dose levels". In: *IEEE Transactions on Medical Imaging* 40.11 (2021), pp. 3065–3076.

[102] Diyu Yang, Shimin Tang, Singanallur V Venkatakrishnan, Mohammad SN Chowdhury, Yuxuan Zhang, Hassina Z Bilheux, Gregery T Buzzard, and Charles A Bouman. "An Edge Alignment-based Orientation Selection Method for Neutron Tomography". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[103] Tianwei Yin, Zihui Wu, He Sun, Adrian V Dalca, Yisong Yue, and Katherine L Bouman. "End-to-End Sequential Sampling and Reconstruction for MRI". In: *Machine Learning for Health*. PMLR. 2021, pp. 261–281.

[104] Mathé T Zeegers, Tristan van Leeuwen, Daniël M Pelt, Sophia Bethany Coban, Robert van Liere, and Kees Joost Batenburg. "A tomographic workflow to enable deep learning for X-ray based foreign object detection". In: *Expert Systems with Applications* 206 (2022), p. 117768.

[105] Mathé Thomas Zeegers. "Spectral imaging and tomographic reconstruction methods for industrial applications". PhD thesis. Leiden University, 2023.

Appendix A

Appendices to Chapter 2

A.1 Parameters used in datasets

Table A.1: Parameters used in datasets

	Scales	**Shifts**
d1) Circles	Radius: 20 $\sim$ 40	Center coordinate: (42,42) $\sim$ (85,85)
d2) Ellipses	Major length: 63 $\sim$ 72, Minor length: 32 $\sim$ 40	Center coordinate: (49,49) $\sim$ (79,79)
d3) Triangles	Right angled sides: 37 $\sim$ 77	Circumscribed circle center: (50,50) $\sim$ (70,70)
d4) Pentagons	Length: 44 $\sim$ 56	Circumscribed circle center: (54,54) $\sim$ (74,74)
d4) Hexagons	Length: 45 $\sim$ 50	Circumscribed circle center: (50,50) $\sim$ (70,70)
d7) Upper ellipse	Major length: 31 $\sim$ 36, Minor length: 8 $\sim$ 13	Center coordinate: (49,49) $\sim$ (79,79)
Lower ellipse	Major length: 39 $\sim$ 41, Minor length: 15 $\sim$ 17	-
d8) Lower ellipse	Major length: 39 $\sim$ 41, Minor length: 15 $\sim$ 17	-

Appendix B

Appendices to Chapter 4

B.1 Continuation advantage function

Notation. $\pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)$ denotes the angle policy, $\pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$ denotes the terminal policy with decision $d \in \{\mathrm{stop}, \mathrm{continue}\}$, and $\pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)$ denotes the transition induced by acquiring angle θ and applying one reconstruction update.

Convention. Whenever d (resp. d', d_k) appears inside $\pi_{\mathrm{ter}}(\cdot \mid \widehat{\boldsymbol{x}})$ without an explicit summation over the decision variable, it denotes the stopping decision. Thus,

$$\pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t) = \pi_{\mathrm{ter}}(\mathrm{stop} \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t),$$

and consequently,

$$1 - \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t) = \pi_{\mathrm{ter}}(\mathrm{continue} \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t).$$

The subscripts in d' and d_k only indicate the associated recursion step; they do not introduce additional free decision variables unless an explicit summation over decisions is written.

The value function decomposes into a stopping branch and a continuation branch:

$$\begin{aligned} V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) = {} & \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)\, \mathrm{PSNR}(\widehat{\boldsymbol{x}}, \bar{\boldsymbol{x}}) \\ & + \Big(1 - \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)\Big)\, V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}). \end{aligned} \tag{B.1}$$

The continuation state-value function is

$$V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) = \sum_{\theta} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)\, Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta). \tag{B.2}$$

The continuation action-value function is

$$Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta) = -b + \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)\, V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}'). \tag{B.3}$$

Continuation advantage. The continuation advantage measures whether the selected

action performs better than the on-policy continuation average at the current state:

$$\begin{aligned} A_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}},\theta) &= Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}},\theta) - \sum_{\theta'} \pi_{\mathrm{a}}(\theta' \mid \widehat{\boldsymbol{x}};\boldsymbol{w}_a)\, Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}},\theta') \\ &= Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}},\theta) - V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) \\ &= -b + \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}},\theta)\, V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}') - V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}) \\ &= -b + \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}},\theta)\Big(\big(1-\pi_{\mathrm{ter}}(d' \mid \widehat{\boldsymbol{x}}';\boldsymbol{w}_t)\big)\, V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}') \\ &\qquad\qquad + \pi_{\mathrm{ter}}(d' \mid \widehat{\boldsymbol{x}}';\boldsymbol{w}_t)\,\mathrm{PSNR}(\widehat{\boldsymbol{x}}',\bar{\boldsymbol{x}})\Big) \\ &\quad - V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}). \end{aligned} \tag{B.4}$$

In practice, since $\pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}},\theta)$ is unknown, we draw a single Monte Carlo sample $\widehat{\boldsymbol{x}}' \sim \pi_t(\cdot \mid \widehat{\boldsymbol{x}},\theta)$ and use the approximation

$$\begin{aligned} A_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}},\theta) \approx\ & -b + \big(1-\pi_{\mathrm{ter}}(d' \mid \widehat{\boldsymbol{x}}';\boldsymbol{w}_t)\big)\, V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}') \\ & + \pi_{\mathrm{ter}}(d' \mid \widehat{\boldsymbol{x}}';\boldsymbol{w}_t)\,\mathrm{PSNR}(\widehat{\boldsymbol{x}}',\bar{\boldsymbol{x}}) \\ & - V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}). \end{aligned} \tag{B.5}$$

B.2 Policy gradient on $\boldsymbol{w}_a$

Notation. $\pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)$ denotes the angle policy, $\pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$ denotes the terminal policy, and $\pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)$ denotes the transition induced by acquiring θ and applying one reconstruction update.

Starting from $J(\boldsymbol{w}_a, \boldsymbol{w}_t) = V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1)$, and using (B.1), we have

$$\begin{aligned}
\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= \nabla_{\boldsymbol{w}_a} V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \\
&= \nabla_{\boldsymbol{w}_a} \Big(\pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\, \mathrm{PSNR}(\widehat{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}) \\
&\qquad + \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big)\, V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \Big) \\
&= \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big)\, \nabla_{\boldsymbol{w}_a} V_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \\
&= \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big) \Bigg(\sum_{\theta} \nabla_{\boldsymbol{w}_a} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a)\, Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1, \theta) \\
&\qquad + \sum_{\theta} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a)\, \nabla_{\boldsymbol{w}_a} Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1, \theta) \Bigg).
\end{aligned} \tag{B.6}$$

Using (B.3), and noting that π_t does not depend on $\boldsymbol{w}_a$, we obtain

$$\nabla_{\boldsymbol{w}_a} Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta) = \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)\, \nabla_{\boldsymbol{w}_a} V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}'). \tag{B.7}$$

Substituting (B.7) into (B.6) yields the recursive form

$$\begin{aligned}
\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) = \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big) \Bigg(& \sum_{\theta} \nabla_{\boldsymbol{w}_a} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a)\, Q_C^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1, \theta) \\
& + \sum_{\theta} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a) \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}_1, \theta)\, \nabla_{\boldsymbol{w}_a} V^{\pi_{\mathrm{a}},\pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}') \Bigg).
\end{aligned} \tag{B.8}$$

Unrolled form. Let $\mu_k(\widehat{\boldsymbol{x}}; \boldsymbol{w}_a, \boldsymbol{w}_t)$ denote the continuation visitation probability of state $\widehat{\boldsymbol{x}}$ at step k, namely the probability of reaching $\widehat{\boldsymbol{x}}$ at step k without terminating earlier. Then

$$\mu_1(\widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a, \boldsymbol{w}_t) = 1,$$

and for $k \geq 1$,

$$\mu_{k+1}(\widehat{\boldsymbol{x}}') = \sum_{\widehat{\boldsymbol{x}}} \mu_k(\widehat{\boldsymbol{x}}) \big(1 - \pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)\big) \sum_{\theta} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)\, \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta). \tag{B.9}$$

By recursively expanding (B.8), we obtain

$$\begin{aligned}\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) = \sum_{k=1}^{T} \sum_{\widehat{\boldsymbol{x}}_k} \mu_k(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a, \boldsymbol{w}_t) \left(1 - \pi_{\text{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)\right) \\ \times \sum_{\theta} \nabla_{\boldsymbol{w}_a} \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\, Q_C^{\pi_{\text{a}}, \pi_{\text{ter}}}(\widehat{\boldsymbol{x}}_k, \theta).\end{aligned} \tag{B.10}$$

Using

$$\nabla_{\boldsymbol{w}_a} \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a) = \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)\, \nabla_{\boldsymbol{w}_a} \log \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a),$$

we may rewrite (B.10) as

$$\begin{aligned}\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) = \sum_{k=1}^{T} \sum_{\widehat{\boldsymbol{x}}_k} \mu_k(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a, \boldsymbol{w}_t) \left(1 - \pi_{\text{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)\right) \\ \times \sum_{\theta} \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\, \nabla_{\boldsymbol{w}_a} \log \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\, Q_C^{\pi_{\text{a}}, \pi_{\text{ter}}}(\widehat{\boldsymbol{x}}_k, \theta).\end{aligned} \tag{B.11}$$

Since

$$\sum_{\theta} \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)\, \nabla_{\boldsymbol{w}_a} \log \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a) = 0,$$

we may subtract the baseline $V_C^{\pi_{\text{a}}, \pi_{\text{ter}}}(\widehat{\boldsymbol{x}}_k)$ without changing the expectation. Hence,

$$\begin{aligned}\nabla_{\boldsymbol{w}_a} J(\boldsymbol{w}_a, \boldsymbol{w}_t) = \sum_{k=1}^{T} \sum_{\widehat{\boldsymbol{x}}_k} \mu_k(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a, \boldsymbol{w}_t) \left(1 - \pi_{\text{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)\right) \\ \times \sum_{\theta} \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\, \nabla_{\boldsymbol{w}_a} \log \pi_{\text{a}}(\theta \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)\, A_C^{\pi_{\text{a}}, \pi_{\text{ter}}}(\widehat{\boldsymbol{x}}_k, \theta).\end{aligned} \tag{B.12}$$

B.3 Policy gradient on $\boldsymbol{w}_t$

Notation. $\pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a)$ denotes the angle policy, $\pi_{\mathrm{ter}}(d \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_t)$ denotes the terminal policy, and $\pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)$ denotes the transition induced by acquiring θ and applying one reconstruction update.

Starting from $J(\boldsymbol{w}_a, \boldsymbol{w}_t) = V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1)$, and using (B.1), we obtain

$$\begin{aligned}
\nabla_{\boldsymbol{w}_t} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= \nabla_{\boldsymbol{w}_t} V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \\
&= \nabla_{\boldsymbol{w}_t} \Big(\pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\, \mathrm{PSNR}(\widehat{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}) \\
&\qquad + \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big)\, V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \Big) \\
&= \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t) \Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \Big) \\
&\quad + \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big)\, \nabla_{\boldsymbol{w}_t} V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1).
\end{aligned} \tag{B.13}$$

Using (B.2) and (B.3), and noting that π_{a} and π_t do not depend on $\boldsymbol{w}_t$, we have

$$\nabla_{\boldsymbol{w}_t} Q_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}, \theta) = \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}, \theta)\, \nabla_{\boldsymbol{w}_t} V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}'). \tag{B.14}$$

Therefore,

$$\begin{aligned}
\nabla_{\boldsymbol{w}_t} J(\boldsymbol{w}_a, \boldsymbol{w}_t) &= \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t) \Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_1) \Big) \\
&\quad + \big(1 - \pi_{\mathrm{ter}}(d_1 \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_t)\big) \sum_{\theta} \pi_{\mathrm{a}}(\theta \mid \widehat{\boldsymbol{x}}_1; \boldsymbol{w}_a) \\
&\qquad \times \sum_{\widehat{\boldsymbol{x}}'} \pi_t(\widehat{\boldsymbol{x}}' \mid \widehat{\boldsymbol{x}}_1, \theta)\, \nabla_{\boldsymbol{w}_t} V^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}').
\end{aligned} \tag{B.15}$$

Using the continuation visitation probability μ_k defined in (B.9), repeated expansion of (B.15) gives

$$\begin{aligned}
\nabla_{\boldsymbol{w}_t} J(\boldsymbol{w}_a, \boldsymbol{w}_t) = \sum_{k=1}^{T} \sum_{\widehat{\boldsymbol{x}}_k} \mu_k(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a, \boldsymbol{w}_t)\, \nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t) \\
\times \Big(\mathrm{PSNR}(\widehat{\boldsymbol{x}}_k, \bar{\boldsymbol{x}}) - V_C^{\pi_{\mathrm{a}}, \pi_{\mathrm{ter}}}(\widehat{\boldsymbol{x}}_k) \Big).
\end{aligned} \tag{B.16}$$

Again, by the convention stated at the beginning of this appendix, d_k inside $\pi_{\mathrm{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)$ denotes the stopping decision. Hence,

$$\nabla_{\boldsymbol{w}_t} \pi_{\mathrm{ter}}(d_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_t)$$

is precisely the gradient of the stopping probability evaluated at the step-k state.

B.4 Additional results

We report supplementary experiments to illustrate robustness across seeds and noise levels (3%, 5%, 7%). Architectures and hyperparameters match Section 5.2.

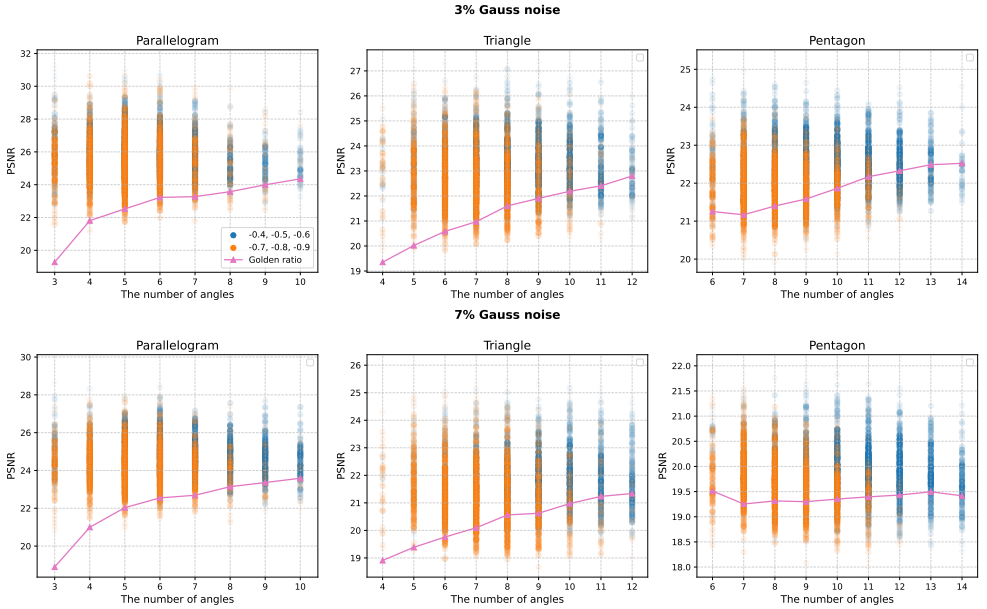


Figure B.1: Simulation results are categorized into parallelogram, triangle, and pentagon shapes to illustrate how the number of angles influences PSNR under various rewards and various noise levels. Data points are color-coded by experimental costs—blue $(-0.4, -0.5, -0.6)$ and orange $(-0.7, -0.8, -0.9)$. Increased transparency indicates data points that are further from the mean. Triangular markers represent the mean PSNR values obtained from the golden ratio policy at each corresponding angle.

The trends mirror those in Figure (4.8). Figure (B.1) encodes ranges of experimental cost b by color: lower costs (blue) encourage acquiring more views, whereas higher costs (orange) penalize additional views and thus lead to fewer selected angles. Across all shapes, the number of acquired angles tends to decrease as b increases, with the separation between cost regimes more pronounced for geometrically complex phantoms that require more views at a given cost level. Moreover, higher noise levels generally drive the policy to select more angles.

B.5 Parameters used in datasets

Table B.1: Parameters used in the datasets. The scale is represented by the radius of the circumscribed circles, and the position is described by the center coordinates (x, y) of these circles.

Shapes	**Scales**	**Shifts**
Parallelogram	Radius: 42 $\sim$ 51	Center (x, y): (110,130) $\sim$ (110,130)
Triangles	Radius: 56 $\sim$ 89	Center (x, y): (110,130) $\sim$ (110,130)
Pentagons	Radius: 56 $\sim$ 89	Center (x, y): (110,130) $\sim$ (110,130)

B.6 Architecture of the Actor–Critic network

The Actor–Critic model consists of a shared convolutional feature extractor followed by three separate heads for policy, value estimation, and terminal state prediction. The complete architecture is detailed in Table (B.2).

Table B.2: Detailed architecture of the Actor–Critic neural network. A shared convolutional encoder extracts features, which are processed by separate heads for policy (actor), value estimation (critic), and terminal prediction.

Layer	Type	Parameters	In Ch.	Out Ch.	Output Size
1	Conv2d	kernel=3, stride=2, padding=1	1	12	120×120
2	GroupNorm	num_groups=4	12	12	same
3	LeakyReLU	negative_slope=0.2	12	12	same
4	MaxPool2d	kernel=2	12	12	60×60
5	Conv2d	kernel=3, padding=1	12	24	same
6	GroupNorm	num_groups=4	24	24	same
7	LeakyReLU	negative_slope=0.2	24	24	same
8	MaxPool2d	kernel=2	24	24	30×30
9	Conv2d	kernel=3, padding=1	24	48	same
10	GroupNorm	num_groups=4	48	48	same
11	LeakyReLU	negative_slope=0.2	48	48	same
12	MaxPool2d	kernel=4	48	48	7×7
13	Flatten	-	-	-	1×2352
Actor Head					
14	Linear	in=1×2352, out=1×180	-	-	1×180
15	Softmax	dim=-1	-	-	1×180
Critic Head					
16	Linear	in=1×2352, out=1×2352	-	-	1×2352
17	ReLU	-	-	-	same
18	Linear	in=1×2352, out=1×1	-	-	1×1
Terminal Head					
19	Linear	in=1×2352, out=1×1	-	-	1×1
20	Sigmoid	-	-	-	1×1

Appendix C

Appendices to Chapter 5

C.1 Power Iteration for Lipschitz Constant Estimation

To ensure stable gradient descent updates in Algorithm 4, the step size is normalized by the Lipschitz constant

$$L = \|\boldsymbol{A}^\top \boldsymbol{W} \boldsymbol{A}\|,$$

which corresponds to the spectral norm of the weighted normal operator. Since exact computation of L is infeasible for large-scale CT systems, we approximate it using the standard *power iteration* method. The procedure is summarized in Algorithm 6.

Algorithm 6 Power Iteration to Estimate Spectral Norm $L = \|\boldsymbol{A}^\top \boldsymbol{W} \boldsymbol{A}\|$

Require: Forward operator $\boldsymbol{A}(\boldsymbol{\theta})$; transpose operator $\boldsymbol{A}^\top(\boldsymbol{\theta})$; weighting operator $\mathcal{W}(\cdot)$; domain size n; maximum iterations T.
Ensure: Spectral norm estimate L.

Initialize: $\boldsymbol{z}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$; normalize $\boldsymbol{z}^{(0)} \leftarrow \boldsymbol{z}^{(0)} / \|\boldsymbol{z}^{(0)}\|$
for $t = 0$ **to** $T - 1$ **do**
  $\boldsymbol{u} \leftarrow \boldsymbol{A}(\boldsymbol{\theta})\, \boldsymbol{z}^{(t)}$ ▷ Forward projection
  $\boldsymbol{v} \leftarrow \mathcal{W}(\boldsymbol{u})$ ▷ Dose-aware weighting
  $\boldsymbol{w} \leftarrow \boldsymbol{A}^\top(\boldsymbol{\theta})\, \boldsymbol{v}$ ▷ Backprojection
  Normalize: $\boldsymbol{z}^{(t+1)} \leftarrow \boldsymbol{w} / \|\boldsymbol{w}\|$
  Update estimate: $L^{(t+1)} \leftarrow \|\boldsymbol{w}\|$
end for
return $L = L^{(T)}$

In practice, only a small number of iterations (e.g., $T \leq 20$) is sufficient to obtain an accurate estimate of the largest eigenvalue, which is then used to set the step size $\alpha = h/L$ in Algorithm 4.

C.2 Additional PPO Training Details

The training loop in Algorithm 5 (main text) summarizes the high-level workflow of PPO in the CT environment. For clarity and readability, minibatch updates, entropy regularization, and advantage estimation are omitted in the main algorithm but are detailed here for reproducibility. We denote the parametric approximation of the value function by $V(\cdot; \boldsymbol{w}_v)$ to distinguish it from the (unknown) true value function $V^\pi(\cdot)$. This convention follows our previous work [97].

Notation.

- $\widehat{\boldsymbol{x}}_k$: reconstructed state at step k,
- θ_k: action (e.g., projection angle or dose) at step k,
- r_k: immediate reward (see Equation 5.22),
- $\pi_a(\cdot \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)$: stochastic policy with parameters $\boldsymbol{w}_a$,
- $V(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v)$: approximate value function (critic) with parameters $\boldsymbol{w}_v$,
- G_k: Monte Carlo return from step k,
- $\widehat{A}_k$: estimated advantage at step k,
- $e_k(\boldsymbol{w}_a) = \dfrac{\pi_a(\theta_k \mid \widehat{\boldsymbol{x}}_k; \boldsymbol{w}_a)}{\pi_a^{\text{old}}(\theta_k \mid \widehat{\boldsymbol{x}}_k)}$: policy probability ratio.

Minibatch Updates. As in standard PPO, parameter updates are performed over multiple epochs using minibatch sampling from the collected trajectory set. Although each episode produces only a small number of transitions (M steps), we construct an effective training batch by aggregating transitions from multiple episodes until a predefined batch size (e.g., 64 samples) is reached. This "batch-size accumulation" strategy stabilizes optimization, ensures sufficient gradient diversity, and enables PPO updates even in environments where each trajectory is short (as in sparse-view CT).

Advantage Estimation. We use generalized advantage estimation (GAE) with discount factor γ and parameter $\lambda \in [0, 1]$. Following the PPO update scheme, the temporal-difference residuals are computed using the value function parameters $\boldsymbol{w}_v^{\text{old}}$, which are frozen during the update epoch:

$$\delta_k = r_k + \gamma V(\widehat{\boldsymbol{x}}_{k+1}; \boldsymbol{w}_v^{\text{old}}) - V(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v^{\text{old}}).$$

The advantages are then estimated as

$$\widehat{A}_k = \sum_{t=0}^{T-k-1} (\gamma\lambda)^t \, \delta_{k+t}, \qquad G_k = \widehat{A}_k + V(\widehat{\boldsymbol{x}}_k; \boldsymbol{w}_v^{\text{old}}).$$

Entropy Regularization. To encourage exploration, we add the policy entropy:

$$\mathcal{H}(\pi_a(\cdot \mid \widehat{\boldsymbol{x}}_k)) = -\sum_{\theta} \pi_a(\theta \mid \widehat{\boldsymbol{x}}_k) \, \log \pi_a(\theta \mid \widehat{\boldsymbol{x}}_k).$$

C.2. Additional PPO Training Details

Total PPO Objective. Combining the actor and critic losses from the main text with entropy regularization, the overall objective is

$$J(\boldsymbol{w}_a, \boldsymbol{w}_v) = \mathbb{E}\big[L^{\text{actor}}(\boldsymbol{w}_a) - c_v\, L^{\text{critic}}(\boldsymbol{w}_v) + c_e\, \mathcal{H}(\pi_a(\cdot \mid \widehat{\boldsymbol{x}}; \boldsymbol{w}_a))\big]\,,$$

where c_v and c_e are positive coefficients weighting the critic and entropy terms, respectively.

Table C.1: Key architectural and algorithmic parameters of the RL pipeline.

Parameter	**Value**
Input image size	256×256
Action-space dimension	180 candidate projection angles
Training seed	1029
Evaluation seeds	$20260526 + r$, where r is the replication index
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clipping parameter ϵ	0.2
PPO update epochs	4
Mini-batch size	64
Timesteps per batch	540
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Training episodes, wedge	1500
Training episodes, foam-defect	2000
Convolutional encoder	
Conv block 1	Conv2D$(1, 12, 3, \text{stride} = 2, \text{padding} = 1)$ + GroupNorm$(4, 12)$ + LeakyReLU(0.2) + MaxPool2D(2)
Conv block 2	Conv2D$(12, 24, 3, \text{padding} = 1)$ + GroupNorm$(4, 24)$ + LeakyReLU(0.2) + MaxPool2D(2)
Conv block 3	Conv2D$(24, 48, 3, \text{padding} = 1)$ + GroupNorm$(4, 48)$ + LeakyReLU(0.2) + MaxPool2D(4)
Actor head	Linear(flattened features, 180) + Softmax
Critic head	Linear(flattened features, flattened features) + ReLU + Linear(flattened features, 1)

List of Publications

Publications included in this thesis

1. Tianyuan Wang, Felix Lucka, and Tristan van Leeuwen. "Sequential Experimental Design for X-Ray CT Using Deep Reinforcement Learning". In: *IEEE Transactions on Computational Imaging* 10 (2024), pp. 953–968. DOI: 10.1109/TCI.2024.3414273

2. Tianyuan Wang, Virginia Florian, Richard Schielein, Christian Kretzer, Stefan Kasperl, Felix Lucka, and Tristan van Leeuwen. "Task-Adaptive Angle Selection for Computed Tomography-Based Defect Detection". In: *Journal of Imaging* 10.9 (2024), p. 208. DOI: 10.3390/jimaging10090208

3. Tianyuan Wang, Felix Lucka, Daniël M. Pelt, K. Joost Batenburg, and Tristan van Leeuwen. "Dynamic Angle Selection in X-Ray CT: A Reinforcement Learning Approach to Optimal Stopping". In: *Applied Mathematics for Modern Challenges* 5 (2025), pp. 36–63. DOI: 10.3934/ammc.2025010

4. Tianyuan Wang, Daniël M Pelt, Felix Lucka, Tristan van Leeuwen, and Kees Joost Batenburg. "Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction". In: *Applied Mathematics in Science and Engineering* 34.1 (2026), p. 2705107. DOI: 10.1080/27690911.2026.2705107

Datasets in this thesis and related publications not included

1. Tianyuan Wang, Felix Lucka, and Tristan van Leeuwen. *X-ray Computed Tomography Case Study: Triangle and Pentagon Datasets with Various Sizes, Scales, and Noise Levels*. Zenodo, Feb. 2025

2. Thomas Schreiber, Christoph Netsch, Sören Eschweiler, Tianyuan Wang, Thomas Storek, Marc Baranski, and Dirk Müller. "Application of data-driven methods for energy system modelling demonstrated on an adaptive cooling supply system". In: *Energy* 230 (2021), p. 120894

Summary

Non-destructive inspection is essential for verifying that products satisfy quality requirements prior to further assembly, processing, or distribution. X-ray computed tomography (CT) is a powerful imaging modality that provides access to the full three-dimensional internal structure of products, enabling reliable identification of internal defects.

In industrial settings, however, inspection is preferably performed inline, without interrupting the production process. Inline inspection is often time-critical and avoids the risks associated with transporting products outside the production line. These requirements make the application of X-ray CT particularly challenging, as conventional CT acquisition typically relies on predefined scanning trajectories and long acquisition times.

From a theoretical perspective, adaptive CT acquisition can be formulated as a sequential Optimal Experimental Design (OED) problem, which requires evaluating expectations over both the prior distribution and the data likelihood at each step. In practice, these distributions are difficult to access or model accurately, and their sequential evaluation is computationally demanding, making efficient angle selection for inline CT a challenging problem.

In this thesis, we address this problem from a new perspective using *reinforcement learning* (RL). RL enables learning adaptive acquisition strategies through offline training, without requiring closed-form models of distributions. Once trained, the resulting policy can be applied efficiently, allowing fast decision-making during inline CT acquisition.

In Chapter 2, we reformulate the sequential OED problem within a reinforcement learning (RL) framework by defining appropriate action spaces, state representations, and reward functions. Using synthetic datasets with known preferential directions, we demonstrate that the RL agent can successfully identify informative projection angles,

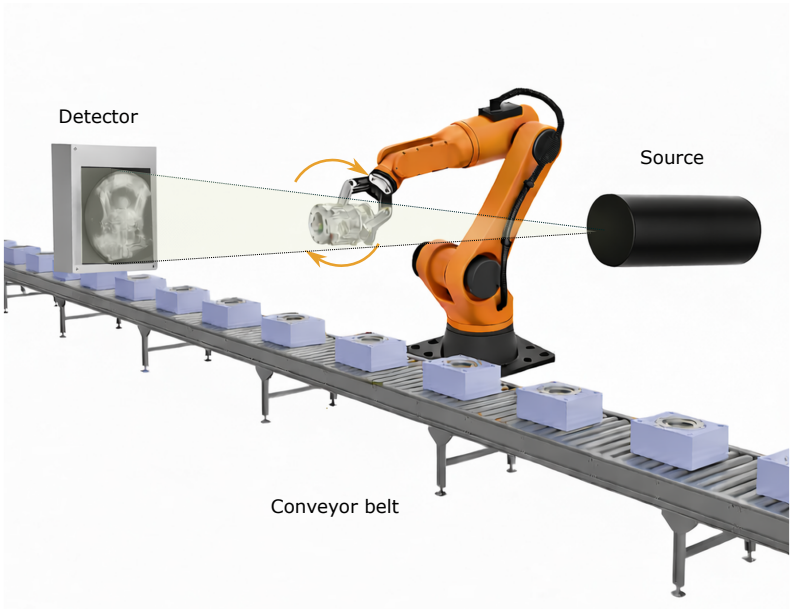


Figure C.1: Inline inspection using X-ray CT. Products are transported along a conveyor belt to the inspection station. At the scanning position, the conveyor belt is temporarily halted, and a robotic manipulator grips and rotates the product to acquire X-ray projections from multiple viewing angles. This controlled rotation enables projection acquisition while maintaining compatibility with an inline inspection workflow. Illustration created by the author; the inline CT system concept was inspired by [33].

which typically correspond to directions tangential to object boundaries or aligned with principal axes.

In Chapter 3, we move closer to practical industrial inspection scenarios by incorporating a downstream task—defect detection—into the design process. The reward function is adapted to reflect task-specific objectives, enabling the RL agent to select projection angles that enhance defect contrast and improve segmentation performance.

In Chapter 4, we extend the framework by introducing a terminal policy that allows the agent to decide when to stop the acquisition process, thereby enabling adaptive selection of the number of projection angles through optimal stopping. In addition, a set of experimental datasets was collected to study the gap between simulation and real-world measurements. The results show that models trained on carefully designed simulated data exhibit reasonable generalization to experimental data.

In Chapter 5, dose allocation is considered jointly with projection angle selection. The reconstruction algorithm is modified to incorporate an angle-dependent weighting

matrix that reflects non-uniform dose levels across projections. Within this setting, the RL agent learns to adaptively select informative angles and simultaneously allocate higher dose to these directions, leading to improved reconstruction quality under a limited total dose budget.

Samenvatting

Niet-destructieve inspectie is essentieel om te verifiëren dat producten voldoen aan kwaliteitsvereisten vóór verdere assemblage, verwerking of distributie. Computertomografie (CT) is een krachtige beeldvormingstechniek die toegang biedt tot de volledige driedimensionale interne structuur van producten en daarmee een betrouwbare identificatie van interne defecten mogelijk maakt.

In industriële omgevingen wordt inspectie echter bij voorkeur inline uitgevoerd, zonder het productieproces te onderbreken. Inline inspectie is vaak tijdkritisch en vermijdt de risico's die gepaard gaan met het transporteren van producten buiten de productielijn. Deze vereisten maken de toepassing van röntgen-CT bijzonder uitdagend, aangezien conventionele CT-acquisitie doorgaans berust op vooraf gedefinieerde scantrajecten en lange acquisitietijden.

Vanuit theoretisch oogpunt kan adaptieve CT-acquisitie worden geformuleerd als een sequentieel Optimal Experimental Design (OED)-probleem, waarbij op elke stap verwachtingen moeten worden geëvalueerd over zowel de a-prioriverdeling als de datalikelihood. In de praktijk zijn deze verdelingen moeilijk toegankelijk of nauwkeurig te modelleren, en is hun sequentiële evaluatie computationeel veeleisend. Hierdoor vormt efficiënte hoekselectie voor inline CT een uitdagend probleem.

In dit proefschrift benaderen wij dit probleem vanuit een nieuw perspectief door gebruik te maken van reinforcement learning (RL). RL maakt het mogelijk adaptieve acquisitiestrategieën te leren via offline training, zonder dat gesloten-vormmodellen van onderliggende verdelingen vereist zijn. Na training kan het resulterende beleid efficiënt worden toegepast, waardoor snelle besluitvorming tijdens inline CT-acquisitie mogelijk wordt.

In Hoofdstuk 2 herformuleren wij het sequentiële OED-probleem binnen een reinforcement learning (RL)-kader door geschikte actieruimten, toestandsrepresentaties en beloningsfuncties te definiëren. Met behulp van synthetische datasets met bekende

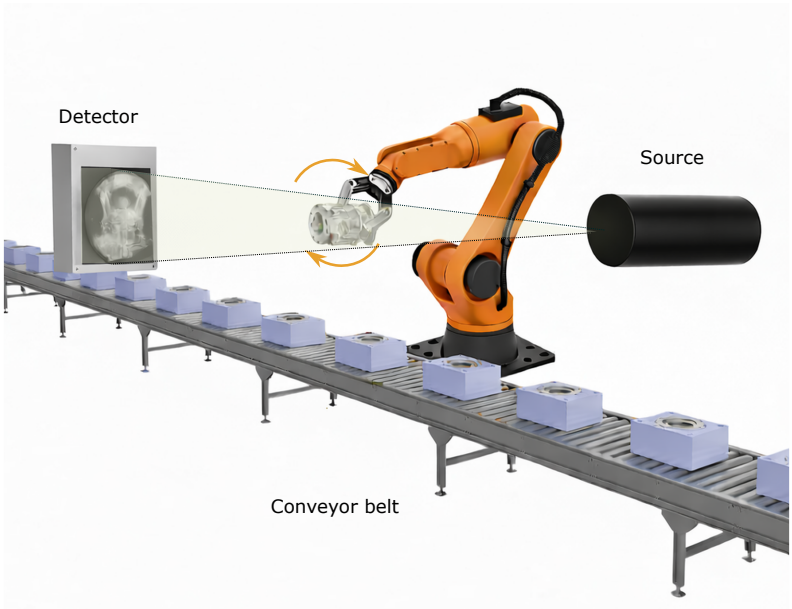


Figure C.2: Inline-inspectie met behulp van röntgen-CT. Producten worden via een transportband naar het inspectiestation vervoerd. Op de scanpositie wordt de transportband tijdelijk stilgezet, waarna een robotmanipulator het product vastgrijpt en roteert om röntgenprojecties vanuit meerdere kijkhoeken te verkrijgen. Deze gecontroleerde rotatie maakt het mogelijk projecties te verkrijgen met behoud van de compatibiliteit met een inline-inspectieproces. Illustratie gemaakt door de auteur; het concept van het inline-CT-systeem is geïnspireerd op [33].

voorkeursrichtingen tonen wij aan dat de RL-agent in staat is informatieve projectiehoeken te identificeren, die doorgaans overeenkomen met richtingen tangentieel aan objectgrenzen of uitgelijnd met de hoofdassen van het object.

In Hoofdstuk 3 benaderen wij realistischere industriële inspectiescenario's door een downstream-taak—defectdetectie—te integreren in het ontwerpproces. De beloningsfunctie wordt aangepast om taak-specifieke doelstellingen te weerspiegelen, waardoor de RL-agent projectiehoeken selecteert die het defectcontrast versterken en de segmentatieprestatie verbeteren.

In Hoofdstuk 4 breiden wij het raamwerk uit door een terminaal beleid te introduceren waarmee de agent kan beslissen wanneer het acquisitieproces moet worden beëindigd. Dit maakt adaptieve selectie van het aantal projectiehoeken mogelijk via optimaal stoppen. Daarnaast is een reeks experimentele datasets verzameld om de kloof tussen simulatie en metingen in de praktijk te bestuderen. De resultaten laten zien dat modellen die zijn getraind op zorgvuldig ontworpen gesimuleerde data een

redelijke generalisatie naar experimentele data vertonen.

In Hoofdstuk 5 wordt dosisallocatie gezamenlijk beschouwd met de selectie van projectiehoeken. Het reconstructie-algoritme wordt aangepast door de introductie van een hoekafhankelijke wegingsmatrix die niet-uniforme dosisniveaus over projecties weerspiegelt. Binnen deze setting leert de RL-agent adaptief informatieve hoeken te selecteren en tegelijkertijd een hogere dosis toe te wijzen aan deze richtingen, wat leidt tot een verbeterde reconstructiekwaliteit onder een beperkt totaal dosisbudget.

Curriculum Vitae

Tianyuan Wang (王天元) was born in 1994 and raised in Huaibei, China. He obtained a Bachelor's degree in Automation from Central South University in Changsha, China, in 2017. Following his undergraduate studies, he moved to Germany to pursue a Master's degree in Computer Engineering at RWTH Aachen University, which he completed in 2020.

In 2021, he began his doctoral research at the Centrum Wiskunde & Informatica (CWI) in Amsterdam under the supervision of Prof. dr. Tristan van Leeuwen and Dr. Felix Lucka, as part of the Marie Skłodowska-Curie Innovative Training Network xCTing. In 2024, he moved to Leiden University to complete the final year of his doctoral research under the supervision of Prof. dr. Joost Batenburg and Dr. Daniël M. Pelt. During his PhD studies, he took a course in scientific conduct and undertook extended research visits to Fraunhofer IIS (Fürth, Germany), KU Leuven (Leuven, Belgium), and Waygate Technologies (Wunstorf, Germany).

Acknowledgments

First, I would like to thank my promotor, Joost Batenburg. Your passion and ability to think outside the box have always inspired me. From you, I also learned the importance of treating others with kindness. I am deeply grateful to Daan Pelt, my daily supervisor at Leiden University during the final year of my PhD, for always being supportive, patient, and helpful. I would also like to thank my second promotor, Tristan van Leeuwen, and my co-promotor, Felix Lucka. You guided me throughout my PhD, taught me invaluable academic skills, and helped shape my research instincts and academic judgment. I am also grateful for the freedom you gave me to explore and pursue my own research ideas.

I would also like to thank all my colleagues at CWI. Because of you, I was fortunate to work in such a welcoming and enjoyable research environment. It was a great pleasure to spend time with Mathé, Richard, Dirk, Marcos, Hamid, Liang, Jinhan, Roozbeh, Ajinkya, Džemila, Maximilian, Adriaan, Poulami, Francien, Ezgi, Alexander, Emma, Robert, Vladyslav, Xinyuan, Yiming and Willem Jan.

I am also grateful to my colleagues and friends from other groups at CWI: Yongqi, Xuemei, Gao, Yuting, Rik, Xinyu, Haohui, Junqiao, Zunchen, Yunhao, Yahui, Sirui and many others. In addition, I would like to thank the support staff at CWI. You are truly the best team: Minnie, Nada, Duda, Doutzen, Paul, Bikkie, Vera, Emil, Remco, and Fiona.

I would also like to thank my friends from the University of Amsterdam: Yixian, Qi, Ruihong, Hongyun, Di, and Xiaoyan. For a while, our Friday dinners became a tradition, and I have many fond memories of the time we spent together.

I am also grateful to my colleagues at LIACS. It was a pleasure to meet and work alongside you during the final year of my PhD: MC, André, Șerban, Jay, and Valentin. I would also like to thank my friends and colleagues from other groups at LIACS: Zhao, Zhong, Lincen, Ying, Yuxuan, Furong, Qi, Kang, Ivan, Hazel, and many others.

I am very glad to have been part of the MSCA ITN 'xCTing' during my PhD. I would like to thank everyone in the xCTing community. We had so much fun travelling and attending conferences together: Javier, Miroslav, Thiago, Pavel, Maryam, Wangduo, Ankit, Negin, Ricardo, Virginia, Zhengrui, Jun, Wim, Stephan, Christian, Natalia, Matthew, Alex, and many others.

I would also like to thank my friends outside work. We spent many enjoyable hours talking, laughing, and simply spending time together: Liang, Anqi, Lu, Wenjing, Xiaorui, Huihao, Kun, Xin, Tianyu, Junxu, and Yang. A special thank-you goes to my two paranymphs, Erik and Jiayang. You brought so much joy and laughter into my life and supported me throughout my PhD.

Finally, I would like to thank my family. Although you are far away, you have always believed in me and been proud of me. Your love, trust, and confidence in me have given me tremendous strength.