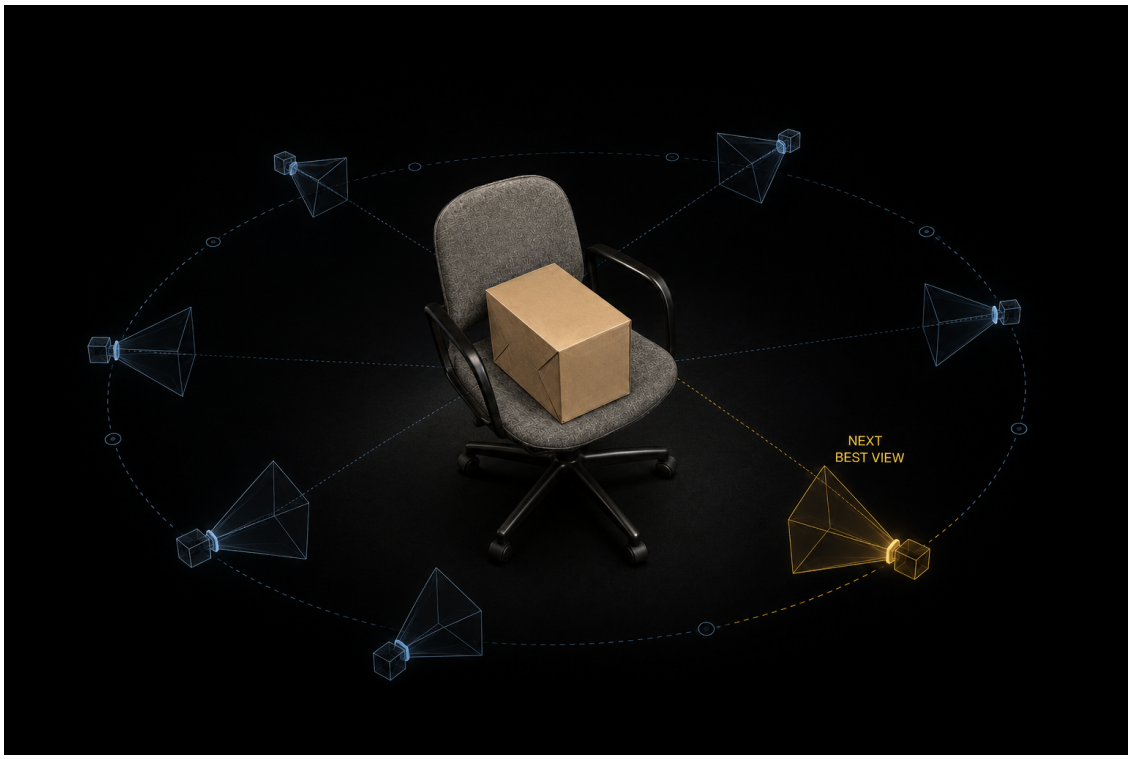
CHALMERS
UNIVERSITY OF TECHNOLOGY



Next-Best-View Planning for Object-Centred 3D Gaussian Splatting

A Controlled Evaluation of Adaptive Image Acquisition and a Quadrotor Proof-of-Concept

Master's thesis in Systems, Control and Mechatronics

GUSTAV ENOCSON, GUSTAV JANSON

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Next-Best-View Planning for Object-Centred 3D Gaussian Splatting

Gustav Enocson, Gustav Janson



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Systems and Control
Chalmers University of Technology
Gothenburg, Sweden 2026

Next-Best-View Planning for Object-Centred 3D Gaussian Splatting
Gustav Enocson, Gustav Janson

© Gustav Enocson, Gustav Janson, 2026.

Supervisor: Knut Åkesson, Electrical Engineering
Company supervisors: Isak Ernstig & Ludwig Friborg, Wirevision
Examiner: Knut Åkesson, Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Division of Systems and Control
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Visualization of multiple camera views and one next best view, around an object.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Next-Best-View Planning for Object-Centred 3D Gaussian Splatting
Gustav Enocson, Gustav Janson
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Accurate Three-Dimensional Gaussian Splatting (3DGS) reconstruction depends strongly on the acquired images — including their viewpoints, overlap, image quality, and estimated camera poses — making image acquisition a key factor in 3DGS reconstruction quality and visual completeness. This thesis investigates an adaptive Next-Best-View (NBV)-guided image acquisition strategy for object-centred 3DGS reconstruction. The aim is to determine whether the NBV strategy improves reconstruction quality, density, and visual completeness for object-centred 3DGS compared to a predefined image acquisition strategy.

A quadrotor-based proof of concept (PoC) was implemented, combining onboard ArUco-based pose estimation and image acquisition with off-board Structure-from-Motion (SfM), 3DGS training, and NBV planning. During testing, several practical limitations — pose-estimation robustness, marker visibility, indoor flight reliability, image quality, and the computational cost of training — prevented the quadrotor from reliably completing the full autonomous workflow. The quadrotor is therefore evaluated as a PoC, while the final comparison between the predefined strategy and the NBV strategy is conducted in a controlled proxy setup using the same reconstruction pipeline and identical 3DGS training settings.

The two strategies are assessed using image registration rate, PSNR, SSIM, and LPIPS on held-out evaluation views, together with reconstruction density, computational cost, and qualitative visual inspection.

The results show that the NBV strategy improves some reconstruction metrics, but not consistently across all comparisons. For the first test object, the NBV strategy achieved higher PSNR and lower LPIPS than the predefined strategy, while the predefined strategy achieved higher SSIM. The NBV-guided reconstructions were also denser, with a greater number of sparse COLMAP points and optimized Gaussians, but required higher computational cost and produced larger models. For the second test object, the predefined strategy outperformed the NBV strategy in most reconstruction metrics, although the NBV strategy again produced denser reconstructions at the cost of higher computational requirements and larger models.

Overall, the results suggest that the NBV strategy has potential as an approach for object-centred 3DGS reconstruction. However, a fully autonomous quadrotor implementation requires more robust pose estimation, flight control, image acquisition, and computational integration before reliable indoor deployment is feasible.

Keywords: 3D Gaussian Splatting, Structure-from-Motion, Next-Best-View, 3D reconstruction, quadrotor, pose estimation, image acquisition

Acknowledgements

We would like to express our sincere gratitude to our academic supervisor and examiner Knut Åkesson at Chalmers University of Technology for his support and guidance throughout this project.

We would also like to thank our company supervisors at Wirevision, Isak Ernstig and Ludwig Friborg, for giving us this opportunity, for providing the hardware and tools necessary to carry out this thesis, and for their support throughout the project.

We would also like to acknowledge the open-source projects and repositories that supported this thesis work. These resources are cited where relevant in the report, and their licenses and contributions are documented in the accompanying public repositories.

Artificial intelligence tools were used in the preparation of this thesis. Claude (Anthropic) and ChatGPT (OpenAI) were used to assist with image generation and writing tasks, including language polishing, reformulating drafts, grammar correction, structural feedback, and peer review of the full thesis. Claude and OpenAI Codex were used to assist with software development tasks throughout the project. AI tools were used over an approximately four-month period during the spring of 2026. All AI-assisted content has been critically reviewed, and the authors take full responsibility for the accuracy, integrity, and conclusions of this work.

Gustav Enocson, Gustav Janson, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis, listed in alphabetical order:

2D	Two-Dimensional
3D	Three-Dimensional
3DGS	Three-Dimensional Gaussian Splatting
AI	Artificial Intelligence
ArUco	Augmented Reality University of Cordoba
CAD	Computer-Aided Design
CPU	Central Processing Unit
CSI	Camera Serial Interface
CUDA	Compute Unified Device Architecture
DDS	Data Distribution Service
DoG	Difference of Gaussians
ESC	Electronic Speed Controller
FOV	Field of View
GPIO	General Purpose Input/Output
GPS	Global Positioning System
GPU	Graphics Processing Unit
IMU	Inertial Measurement Unit
LiPo	Lithium Polymer
LPIPS	Learned Perceptual Image Patch Similarity
MI	Mutual Information
NBV	Next-Best-View
NeRF	Neural Radiance Field
OS	Operating System
PDB	Power Distribution Board

PnP	Perspective-n-Point
PoC	Proof-of-Concept
PSNR	Peak Signal-to-Noise Ratio
PWM	Pulse Width Modulation
QGC	QGroundControl
RGB	Red, Green, Blue
ROI	Region of Interest
ROS	Robot Operating System
RPi	Raspberry Pi
RQ	Research Question
RX	Receiver
SDG	Sustainable Development Goal
SfM	Structure-from-Motion
SH	Spherical Harmonics
SIFT	Scale-Invariant Feature Transform
SSIM	Structural Similarity Index Measure
SSH	Secure Shell
TELEM	Telemetry
TX	Transmitter
UART	Universal Asynchronous Receiver-Transmitter
UAV	Unmanned Aerial Vehicle
UWB	Ultra-Wideband
VRAM	Video Random Access Memory
XRCE	Extremely Resource Constrained Environments

Nomenclature

Below is the nomenclature of indices, coordinate frames, parameters, and variables that have been used throughout this thesis.

Indices

i	Index for Gaussian primitives
j	Index for pixels or accumulation order
l	Degree index in the spherical harmonics representation
m	Order index in the spherical harmonics representation
p	Pixel index
t	Iteration index in optimization

Coordinate Frames

$\{W\}$	World frame defined by the reference ArUco marker
$\{M\}$	ArUco marker frame
$\{C_t\}$	Tracking-camera frame
$\{C_c\}$	Capture-camera frame

Parameters

η	Learning rate or step size used in optimization
α_i	Opacity of Gaussian i
σ	Scale parameter in the Gaussian scale-space representation
k	Scale factor between neighbouring levels in scale space
f_x, f_y	Camera focal lengths in pixels
c_x, c_y	Camera principal point coordinates

s	Camera skew parameter
s_x, s_y, s_z	Scale values of a Gaussian along its principal axes
λ	Isotropic scale factor for an initially spherical Gaussian
w_1, w_2, w_3	Weights in the multi-metric NBV utility function

Variables and Matrices

G_i	Gaussian primitive i
μ_i	Mean position of Gaussian i
Σ_i	Covariance matrix of Gaussian i
c_i	Colour of Gaussian i
$c(v)$	View-dependent colour in viewing direction v
$c_{l,m}$	Learned spherical harmonics coefficient of degree l and order m
$Y_{l,m}(v)$	Spherical harmonics basis function evaluated in direction v
C	Rendered pixel colour
C_p	Rendered colour at pixel p
$I_{\text{obs},p}$	Observed colour at pixel p in the acquired image
T_i	Transmittance up to Gaussian i
v	Viewing direction vector
K	Camera intrinsic matrix
K_t	Tracking-camera intrinsic matrix
K_c	Capture-camera intrinsic matrix
R	Rotation matrix
S	Scale matrix of a Gaussian
R_{AB}	Rotation from frame $\{B\}$ to frame $\{A\}$
t_{AB}	Translation of frame $\{B\}$ expressed in frame $\{A\}$
T_{AB}	Homogeneous transformation matrix from frame $\{B\}$ to frame $\{A\}$
T_{WC_t}	Pose of the tracking camera in the world frame
T_{C_tW}	Transform from the world frame to the tracking-camera frame
$T_{C_tC_c}$	Transform from the capture-camera frame to the tracking-camera frame
T_{WC_c}	Pose of the capture camera in the world frame
p_A	3D point expressed in frame $\{A\}$
p_B	3D point expressed in frame $\{B\}$

$I(x, y)$	Image intensity at image coordinates (x, y)
$G(x, y, \sigma)$	Gaussian kernel at scale σ
$L(x, y, \sigma)$	Gaussian-blurred image at scale σ
$D(x, y, \sigma)$	Difference-of-Gaussians representation
L	Reconstruction loss function
θ_i	Parameter vector of Gaussian i, $\theta_i = (\mu_i, \Sigma_i, c_i, \alpha_i)$
q	Candidate camera viewpoint or pose
$U(q)$	Utility of candidate viewpoint q
$U_{\text{multi}}(q)$	Multi-metric uncertainty utility of candidate viewpoint q
$U_{\text{opacity}}(q)$	Utility term for low-opacity regions visible from q
$U_{\text{coverage}}(q)$	Utility term for newly visible or poorly covered regions
$U_{\text{surface}}(q)$	Utility term for surface uncertainty
$U_{\text{MI}}(q)$	Mutual-information score of candidate viewpoint q
$H(\theta)$	Entropy of the current probabilistic scene representation
θ	Probabilistic scene representation used in the information-gain formulation
z_v	Hypothetical observation from viewpoint q
$\nabla_v U(q)$	Gradient of the utility function with respect to the camera pose
$S(q)$	Final NBV score of candidate viewpoint q
$d_{\text{existing}}(q)$	Discount for similarity to existing camera poses
$d_{\text{tried}}(q)$	Discount for similarity to previously suggested NBV poses
$b_{\text{coverage}}(q)$	Elevation-coverage bonus for candidate viewpoint q
$h(q)$	Height prior for candidate viewpoint q

Contents

List of Acronyms ix

Nomenclature xii

List of Figures xxi

List of Tables xxiii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Purpose . . . 2
- 1.3 Objective . . . 2
  - 1.3.1 Research Questions . . . 3
- 1.4 Scope . . . 3
- 1.5 Ethical and sustainability aspects . . . 4

2 Problem Definition and Overall Approach 5
- 2.1 Overall Problem . . . 5
- 2.2 Decomposition into three subproblems . . . 6
- 2.3 Overall approach . . . 6
- 2.4 Motivation for key design choices . . . 7
  - 2.4.1 Use of quadrotor as image acquisition platform . . . 7
  - 2.4.2 Use of ArUco markers for pose estimation . . . 7
  - 2.4.3 Use of predefined and adaptive image acquisition strategies . . . 8
  - 2.4.4 Use of COLMAP and 3DGS for 3D reconstruction . . . 8
  - 2.4.5 Use of NBV planner . . . 9
- 2.5 System Overview . . . 9

3 Pose Estimation 11
- 3.1 Problem . . . 11
- 3.2 Theory . . . 11
  - 3.2.1 Camera model and coordinate frames . . . 12
    - 3.2.1.1 World frame definition . . . 12
    - 3.2.1.2 Quadrotor and camera frames . . . 12
    - 3.2.1.3 Pose composition . . . 13
    - 3.2.1.4 Camera intrinsics and calibration . . . 14
  - 3.2.2 Pose estimation methods . . . 14

3.2.2.1 ArUco-based pose estimation . . . 14
3.2.2.2 Structure from Motion . . . 15
3.2.2.3 Georeferencing . . . 15
3.3 Method . . . 16
3.3.1 Experimental setup . . . 16
3.3.2 ArUco-based pose estimation experiment . . . 18
3.3.3 SfM-based pose estimation experiment . . . 18
3.3.4 Combined ArUco + SfM experiment . . . 18
3.4 Validation . . . 18
3.4.1 ArUco-based pose estimation . . . 19
3.4.2 SfM-based pose estimation . . . 19
3.4.3 Combined ArUco + SfM-based pose estimation . . . 20

4 3D Reconstruction **21**
4.1 Problem . . . 21
4.2 Theory of 3D Gaussian Splatting . . . 21
4.3 Method . . . 23
4.3.1 Sparse point cloud generation . . . 23
4.3.2 3DGS training . . . 24
4.3.3 3DGS model post-processing . . . 25
4.4 Validation . . . 26
4.4.1 Sparse point cloud . . . 26
4.4.2 3DGS reconstruction . . . 26

5 Next-Best-View Planning **29**
5.1 Problem . . . 29
5.2 Theory . . . 29
5.2.1 Next-Best-View planning . . . 29
5.2.2 Reconstruction representations for NBV . . . 30
5.2.3 3DGS for NBV . . . 30
5.3 Method . . . 31
5.3.1 NBV planner overview . . . 31
5.3.2 Reliability update . . . 32
5.3.3 Candidate viewpoint generation . . . 32
5.3.4 Candidate evaluation and scoring . . . 33
5.3.5 Candidate ranking and output . . . 33
5.4 Validation . . . 33

6 System Architecture **35**
6.1 Hardware setup . . . 35
6.1.1 Camera setup . . . 36
6.2 Software architecture . . . 37
6.3 Flight procedure . . . 38
6.4 Implementation status . . . 40
6.5 Summary . . . 41

7 Evaluation and Results **43**

7.1 Evaluation objective . . . 43
7.2 Evaluation setup . . . 44
7.3 Evaluation metrics . . . 45
7.3.1 Image registration and reconstruction information . . . 47
7.3.2 Image-based reconstruction quality . . . 47
7.3.3 Computational and practical cost . . . 48
7.4 Visual comparison . . . 48
7.5 Results . . . 48
7.5.1 Image registration and reconstruction information . . . 48
7.5.2 Image-based reconstruction quality . . . 49
7.5.3 Computational and practical cost . . . 52
7.5.4 Visual comparison . . . 53
7.5.4.1 Renders of Office Chair with Box . . . 53
7.5.4.2 Quadrotor . . . 54
7.6 Quadrotor PoC: practical limitations . . . 55
7.7 Summary of evaluation . . . 55

8 Discussion **57**
8.1 Changing the image capturing system . . . 57
8.2 Discussion of Research Questions . . . 58
8.3 Limitations of the evaluation . . . 61
8.4 Future Work . . . 63
8.5 Conclusion . . . 64

A Appendix **I**

List of Figures

3.1 Definition of the world frame based on the reference ArUco marker. . 12
3.2 Pixel neighbours in scale-space. Created by the author with assistance from ChatGPT, OpenAI, on 4 June 2026. 16
3.3 CAD model of the test rig. 17

4.1 Sparse point cloud reconstruction with estimated camera poses. . . . 26
4.2 3DGS reconstruction results. 27

5.1 Visualization of the initial NBV planning result. 34

6.1 PX4 Development Kit X500 v2 from Holybro. 36
6.2 Pixhawk 6C. 36
6.3 Flowchart of the quadrotor hardware. 37
6.4 Client-agent communication between the Pixhawk and RPi. 38
6.5 Flowcharts of the two flight procedures. 39

7.1 Evolution of PSNR for the predefined and NBV strategies. 50
7.2 Evolution of SSIM for the predefined and NBV strategies. 51
7.3 Evolution of LPIPS for the predefined and NBV strategies. 51
7.4 Rendered images of each 3DGS model from the four different strategies, captured from the same above and side view in each 3DGS model. 53
7.5 Rendered images of each 3DGS model from the four different strategies, captured from the same back view in each 3DGS model. 54
7.6 Rendered images of each 3DGS model from the four different strategies. 55

List of Tables

3.1 Translation and rotation errors for ArUco-based pose estimation. . . . 19
3.2 Translation and rotation errors for SfM-based pose estimation. 19
3.3 Translation and rotation errors for combined ArUco + SfM-based pose estimation. 20

6.1 Implementation status of the quadrotor PoC system. 41

7.1 Evaluation cases for the predefined and NBV-guided image acquisition strategies. 45
7.2 Training-image counts and cumulative training iterations used in the staged reconstruction comparison. 46
7.3 Fixed reconstruction and evaluation settings used for both strategies. 46
7.4 Image registration and reconstruction density results. 49
7.5 Stage-wise image-based reconstruction quality results. Each pair of rows compares the predefined and NBV strategies using the same number of training images. Bold values indicate the better result within each predefined/NBV pair. 50
7.6 Final image-based reconstruction quality results at 36 training images. Bold values indicate the better result within each predefined/NBV pair. 51
7.7 Computational reconstruction results. 52
7.8 Practical limitations of the quadrotor PoC system. 56

1
Introduction

This thesis investigates adaptive image acquisition for object-centred 3D Gaussian Splatting (3DGS) reconstruction. A quadrotor-based proof-of-concept (PoC) system is implemented to study how onboard pose estimation and image capture can be integrated with off-board Structure-from-Motion (SfM), 3DGS reconstruction, and Next-Best-View (NBV) planning. Since the implemented quadrotor platform was not sufficiently reliable for repeatable autonomous indoor acquisition, the final comparison between predefined and NBV-based image acquisition is performed in a controlled proxy setup.
The central problem is therefore not only how to reconstruct an object from images, but also how to acquire images that are useful for reconstruction. Predefined image acquisition strategies are simple and repeatable, but they do not adapt to the current state of the 3DGS reconstruction. As a result, they may collect redundant images of already well-observed regions while leaving other parts of the object insufficiently covered. An adaptive image acquisition strategy based on an NBV planner may address this limitation by selecting new viewpoints based on the current state of the 3DGS model.
This thesis therefore separates two questions: whether NBV-based viewpoint selection can improve object-centred 3DGS reconstruction in a controlled evaluation setup, and what practical limitations prevent the implemented quadrotor PoC system from completing the same workflow autonomously. The work is motivated both by research interest in reconstruction-guided image acquisition and by an industrial need for efficient generation of accurate 3DGS models. This chapter presents the background, purpose, objective, scope, and ethical and sustainability aspects of the thesis.

1.1 Background

Image-based 3D reconstruction is an important tool in applications where physical objects or environments need to be represented digitally. In industrial contexts, accurate 3D representations can support documentation, inspection, quality control, and digital-twin workflows [1]. In research and robotics, reconstruction also plays an important role in perception, mapping, and scene understanding.
A central challenge in 3DGS reconstruction is that the final 3DGS model depends not only on the reconstruction method, but also on how the image data is collected. If important object surfaces are not observed from sufficiently informative viewpoints, the resulting 3DGS model may become incomplete or inaccurate. Image acquisition

is therefore an important part of the reconstruction problem rather than only a data collection step.
Quadrotor-based image acquisition is relevant because it enables automated and flexible access to viewpoints around an object. This is particularly useful when manual image acquisition is inefficient, difficult to repeat consistently, or unsuitable for larger objects. However, quadrotor-based image acquisition also introduces practical challenges related to motion, sensing uncertainty, pose estimation, and viewpoint planning. These challenges motivate the investigation of image acquisition strategies that can support reliable and efficient 3DGS reconstruction.
NBV planning provides one approach for making image acquisition adaptive. Instead of following a fully predefined acquisition path, an NBV strategy selects future viewpoints based on what has already been observed. In 3D reconstruction, partial reconstructions can therefore be used to guide view planning, with the aim of improving surface coverage and reducing redundant image acquisition [2].
Recent work has also investigated 3DGS-based representations for active reconstruction. GauSS-MI [3] proposes an online active reconstruction framework in which Gaussian reliability, estimated using a mutual-information (MI) formulation and a modified rasterizer, is used to select the next camera viewpoint during continuous data collection. This demonstrates that 3DGS-based MI scoring can be useful for active reconstruction.
However, the setting considered in this thesis differs from such online active reconstruction frameworks. In this work, image acquisition and planning are decoupled: the quadrotor acquires images, lands, processes data off-board, and then returns to acquire additional images. The candidate viewpoints are also constrained to a predefined physical acquisition geometry. These constraints motivate the investigation of whether MI-based NBV planning can be adapted to an offline, discrete-candidate setting for object-centred 3DGS reconstruction.

1.2 Purpose

The purpose of this thesis is to investigate whether adaptive image acquisition using NBV planning can improve object-centred 3DGS reconstruction compared with a predefined acquisition strategy. A second purpose is to assess the feasibility of integrating pose estimation, image capture, reconstruction, and NBV planning into a quadrotor-based PoC platform for future fully-autonomous acquisition.
The work is also motivated by an industrial need to generate accurate 3DGS models of products from images acquired together with pose information. In this context, the thesis serves as an initial PoC study of a future system for fully-autonomous quadrotor-based image acquisition.

1.3 Objective

The objective of this thesis is to investigate adaptive image acquisition for object-centred 3DGS reconstruction using NBV planning. The work has two connected aims: first, to implement and test a quadrotor-based PoC system for camera-based

pose estimation and image acquisition; and second, to evaluate whether an adaptive NBV-based acquisition strategy can improve reconstruction quality or acquisition efficiency compared with a predefined strategy.
The quadrotor platform is developed as a PoC system to study the practical requirements of autonomous image acquisition. The system combines ArUco-based pose estimation, image capture, SfM, 3DGS reconstruction, and NBV planning. Its purpose is to assess the feasibility of integrating these components into a future autonomous acquisition workflow.
Because the implemented quadrotor system was not sufficiently robust for repeatable autonomous indoor acquisition, the final comparison between the predefined and adaptive strategies is performed using a controlled proxy setup. This allows the acquisition strategies to be compared under repeatable conditions while preserving the main reconstruction and NBV-planning workflow.
The evaluation focuses on whether the adaptive strategy can improve reconstruction quality and visual completeness when using the same number of images, or whether comparable reconstruction quality can be achieved with fewer images or less acquisition effort. The thesis also identifies the main practical limitations that must be addressed before the same workflow can be completed reliably using an autonomous quadrotor.

1.3.1 Research Questions

The thesis is guided by the following research questions:

- **RQ1**: How does image-based reconstruction quality and visual completeness compare between an adaptive NBV-based and a predefined image acquisition strategy when using the same number of images?
- **RQ2**: How does the number of images required to reach comparable image-based reconstruction quality and visual completeness differ between an adaptive NBV-based and a predefined image acquisition strategy in a controlled proxy setup?
- **RQ3**: What practical limitations prevent the implemented quadrotor PoC system from completing the same acquisition and reconstruction workflow autonomously in an indoor environment?

1.4 Scope

This thesis is scoped as a PoC study of adaptive image acquisition for object-centred 3DGS reconstruction. The work focuses on integrating pose estimation, SfM, 3DGS reconstruction, and an NBV planner into a common workflow for comparing predefined and adaptive image acquisition strategies.
The quadrotor is implemented and tested as a PoC platform for image acquisition using an ArUco-based tracking camera and a separate capture camera. The scope includes evaluating the feasibility of this setup and identifying practical limitations that affect reliable indoor operation.
The final comparison between the predefined strategy and the adaptive NBV-based strategy is performed in a controlled proxy setup. This allows the acquisition strate-

gies to be evaluated under repeatable conditions using the same reconstruction pipeline, while the quadrotor system is treated as the intended PoC platform for future autonomous acquisition. Therefore, the final evaluation should be interpreted as an evaluation of the acquisition strategies and reconstruction pipeline, not as a full validation of autonomous quadrotor acquisition.
The experiments are limited to static objects in a controlled indoor environment. Outdoor operation, large-scale scenes, onboard 3DGS training, fully autonomous real-time operation, and product-level robustness are outside the scope of this thesis.

1.5 Ethical and sustainability aspects

This project includes ethical and sustainability considerations because quadrotor-based 3DGS reconstruction systems could be used in industrial, commercial, or public environments. The possible consequences depend on how such systems are developed and deployed. In this thesis, the system is evaluated only as a controlled PoC, but the broader implications are still relevant to consider.

Positive Consequences

- **Efficiency and resource optimization:** Adaptive image acquisition may reduce redundant image acquisition, which could reduce flight time, energy use, and processing time.
- **Improved digital documentation and quality control:** Accurate 3DGS models can support inspection, digital twins, remote monitoring, and documentation of physical objects.
- **Contribution to UN Sustainable Development Goals (SDGs):** The project is related to Goal 9 (Industry, Innovation, and Infrastructure), Goal 12 (Responsible Consumption and Production), and Goal 13 (Climate Action), particularly through its potential to support more efficient industrial inspection and documentation workflows.

Negative Consequences

- **Safety and privacy:** Quadrotor-based image acquisition systems may create safety and privacy risks if they are operated in uncontrolled environments or used to collect visual data without appropriate restrictions.
- **Job displacement:** Automating image acquisition and inspection-related tasks could reduce the need for some manual work in similar settings.
- **Environmental impact:** Quadrotor operation has a minor environmental impact through energy use, noise, and possible disturbance of humans or wildlife.
- **Misuse of autonomous reconstruction systems:** Similar systems could be misused for surveillance, planning, or navigation in harmful applications. This risk is not specific to this thesis implementation, but it is relevant when developing autonomous perception and reconstruction technologies.

2

Problem Definition and Overall Approach

This chapter defines the technical problem addressed in the thesis and explains how it is divided into three connected subproblems. It also presents the overall workflow and motivates the main design choices used in the PoC system. The purpose of the chapter is to connect the research questions from Chapter 1 to the technical chapters that follow.

2.1 Overall Problem

The overall problem addressed in this thesis is how image acquisition can be guided to improve object-centred 3DGS reconstruction. More specifically, the thesis investigates whether an adaptive image acquisition strategy based on NBV planning can improve reconstruction quality and visual completeness compared with a predefined image acquisition strategy, or achieve comparable reconstruction results with fewer acquired images.

This problem is challenging because the usefulness of each acquired image depends on several coupled factors, including image quality, camera pose estimation, viewpoint selection, and overlap between images. In a predefined image acquisition strategy, these factors are determined before the reconstruction is available. In an adaptive image acquisition strategy, the current 3DGS model is instead used to guide future image acquisition, which creates a direct dependency between reconstruction quality and viewpoint planning.

The 3DGS model therefore has two roles in this thesis. It is both the final reconstruction output and the intermediate representation used by the NBV planner. If the current 3DGS model is incomplete or inaccurate, the NBV planner may select viewpoints that do not improve the reconstruction. The problem is therefore highly coupled: pose estimation, 3DGS reconstruction, and NBV planning must function together for adaptive image acquisition to be meaningful.

In addition to evaluating the adaptive acquisition strategy, this thesis implements and tests a quadrotor-based PoC system for object-centred image acquisition. The quadrotor platform is used to investigate the practical requirements and limitations of autonomous indoor acquisition, while the final comparison between predefined and adaptive acquisition strategies is performed in a controlled proxy setup. This separates the evaluation of the acquisition strategy from the practical limitations of the quadrotor platform. This distinction is important because the controlled proxy

setup is used to answer the acquisition-strategy questions, while the quadrotor PoC system is used to identify practical limitations of autonomous acquisition.

2.2 Decomposition into three subproblems

The overall problem is divided into three subproblems: pose estimation, 3D reconstruction, and NBV planning. This decomposition reflects the dependency chain in the proposed workflow, where pose estimation enables 3DGS reconstruction, the 3DGS model supports NBV planning, and the NBV planner guides continued image acquisition.
The pose estimation subproblem provides the camera poses required to relate all acquired images to a common world frame. The input to this subproblem consists of camera observations and known reference information, such as ArUco marker geometry or image correspondences from SfM. The output is a set of capture-camera poses expressed in the ArUco-defined world frame $\{W\}$. These poses are required both for 3DGS reconstruction and for relating the reconstructed object to the physical image acquisition setup.
The 3D reconstruction subproblem transforms acquired images and associated camera poses into a 3DGS model of the observed object. In this thesis, the 3DGS model is not only evaluated as a final output, but also used as an intermediate representation for adaptive image acquisition. It must therefore provide enough spatial information for the NBV planner to identify which parts of the object are already well covered and which parts may benefit from additional viewpoints.
The NBV planning subproblem uses the current 3DGS model to select new viewpoints for image acquisition. The input is the intermediate 3DGS model, and the output is one or more candidate viewpoints expected to improve the reconstruction. This subproblem makes the workflow adaptive because future image acquisition depends on the current state of the 3DGS model rather than on a fully predefined viewpoint sequence.
Together, these subproblems define the technical workflow evaluated in the thesis. Pose estimation enables reconstruction, reconstruction provides the intermediate 3DGS representation, and NBV planning uses this representation to guide continued image acquisition.

2.3 Overall approach

The overall approach is to compare two image acquisition strategies for 3DGS reconstruction. The first is a predefined strategy, where the viewpoints are selected before reconstruction. The second is an adaptive strategy, where an NBV planner selects new viewpoints based on the current 3DGS model. In both cases, the acquired images and associated camera poses are used to reconstruct the object.
The intended autonomous workflow begins with a quadrotor acquiring an initial image set, where camera poses are estimated and associated with the captured images. The acquired images and poses are processed using COLMAP, aligned to the ArUco-defined world frame, and used to train an initial 3DGS model. This model

is then used by the NBV planner to evaluate candidate viewpoints and propose additional views.
In the intended autonomous workflow, the quadrotor would acquire images from the NBV-proposed viewpoints, after which reconstruction and planning would be repeated. During implementation, however, the quadrotor system was not sufficiently reliable for repeatable autonomous acquisition. The final strategy comparison is therefore performed using a controlled proxy setup, which preserves the reconstruction and NBV-planning workflow while avoiding flight instability and acquisition repeatability as confounding factors.
The comparison is used to evaluate whether NBV-based acquisition improves image-based reconstruction quality, visual completeness, or acquisition efficiency. The quadrotor implementation is evaluated separately as a PoC system to identify limitations that must be resolved before the workflow can be completed autonomously.

2.4 Motivation for key design choices

The following subsections motivate key design choices made for this thesis.

2.4.1 Use of quadrotor as image acquisition platform

A quadrotor was selected as the intended image acquisition platform because it provides flexible access to viewpoints around an object and can in principle support repeatable and automated viewpoint acquisition. This is relevant for future industrial reconstruction workflows in which manual image capture may be inefficient, difficult to standardise, or unsuitable for larger objects.
In this thesis, however, the quadrotor is treated as a PoC platform rather than as the platform used for the final quantitative comparison. The system was implemented and tested to investigate the practical requirements of autonomous indoor acquisition, including flight execution, camera integration, ArUco-based pose estimation, and integration with the reconstruction and NBV-planning pipeline.
This design choice also introduced practical complexity compared with a controlled manual or rig-based acquisition setup. These limitations made the platform valuable for studying the intended autonomous workflow, but less suitable for the repeatable final comparison of acquisition strategies.

2.4.2 Use of ArUco markers for pose estimation

ArUco markers are used to support pose estimation because they provide a practical, low-cost, and well-documented fiducial marker solution for camera pose estimation and robot localization [4]. They are suitable for a PoC setting because they provide a simple and repeatable way to obtain metric pose estimates in a controlled environment. Since the physical marker geometry is known, ArUco markers also provide metric pose estimates that can be used to establish a consistent world frame.
The main limitation is that ArUco-based pose estimation depends on marker visibility and image quality. The marker layout must be known in advance, and the pose estimate can be affected by distance, viewing angle, resolution, lighting, and motion

blur. These limitations are acceptable in this thesis because the experiments are performed in a controlled environment and because the main focus is on evaluating the adaptive image acquisition workflow rather than developing a robust localization system for all operating conditions.

2.4.3 Use of predefined and adaptive image acquisition strategies

A predefined image acquisition strategy is included as a baseline because it is simple, repeatable, and easy to interpret. The viewpoint sequence can be planned before image acquisition, which makes the approach operationally straightforward. However, a predefined strategy does not adapt to the current state of the 3DGS model and may therefore acquire redundant images of already well-covered regions while leaving other parts of the object insufficiently observed [5]. The specific predefined geometry used in this thesis was chosen to reflect how a non-expert drone operator would typically program a simple object survey: a repeatable path based on the approximate shape of the object, planned without prior knowledge of the reconstruction state.

An adaptive image acquisition strategy is included because the value of a new image depends on what has already been observed. By using the current 3DGS model to guide image acquisition, the system can prioritize viewpoints that are expected to add useful information to the reconstruction. This creates the possibility of improving 3DGS model quality and surface coverage, or achieving comparable reconstruction performance with fewer images or less image acquisition effort [5].

This comparison makes it possible to evaluate whether the additional complexity of adaptive image acquisition is justified.

2.4.4 Use of COLMAP and 3DGS for 3D reconstruction

The reconstruction pipeline combines sparse reconstruction using Structure from Motion (SfM), implemented through COLMAP, with 3DGS as the main reconstruction method. COLMAP is used because it provides a well-established foundation for estimating camera poses and generating a sparse point cloud from multiple overlapping images [6]. This sparse reconstruction provides a practical initialization for the later 3DGS reconstruction stage.

3DGS is used because it provides a dense and visually informative representation of the observed object, while also supporting efficient rendering [7]. In contrast to a sparse point cloud, a 3DGS model can represent the object in a way that is more useful for evaluating reconstruction quality and for supporting adaptive viewpoint planning.

This choice also introduces computational complexity. Since 3DGS training requires GPU resources beyond the implemented onboard platform, COLMAP, 3DGS training, and NBV planning are performed off-board. The focus is therefore on evaluating the adaptive acquisition workflow rather than demonstrating onboard real-time reconstruction.

2.4.5 Use of NBV planner

An NBV planner is used because adaptive image acquisition requires a method for selecting new viewpoints based on the current reconstruction. In contrast to a predefined strategy, where all viewpoints are selected before reconstruction, the NBV planner evaluates candidate viewpoints using the current 3DGS model and selects views expected to improve the reconstruction.
The main limitation is that the planner depends on the quality of the intermediate 3DGS model and on the chosen viewpoint-evaluation criterion. If the intermediate reconstruction is inaccurate or incomplete, the selected viewpoints may not improve the final model.

2.5 System Overview

The implemented PoC is built around a quadrotor carrying two cameras, a Pixhawk 6C flight controller, and an onboard Raspberry Pi (RPi) 5. A downward-facing tracking camera is used for ArUco-based pose estimation, while a servo-mounted capture camera is used for image acquisition. The RPi communicates with the Pixhawk 6C and handles marker detection, pose estimation, and image capture.
COLMAP, 3DGS reconstruction, and NBV planning are performed off-board because 3DGS training requires GPU resources that are not available onboard the implemented platform. The system is therefore designed as a distributed PoC workflow combining onboard acquisition with off-board reconstruction and viewpoint planning. Off-board processing was chosen for two connected reasons. First, 3DGS training required GPU resources that are not available on the onboard RPI. Second, and by design, this PoC focuses on evaluating the adaptive acquisition and planning workflow rather than demonstrating onboard real-time reconstruction. A future system with sufficient onboard compute, or with fast off-board connection, could in principle close the loop more tightly; that remains outside the scope of this thesis.
The detailed hardware and software architecture of the quadrotor PoC is presented in Chapter 6, while the controlled evaluation setup used for the final strategy comparison is described in Chapter 7.

3
Pose Estimation

This chapter addresses the pose estimation subproblem of the project. The purpose is to define the coordinate systems and pose formulations needed to express camera poses in a common world frame, and to validate whether the selected pose estimation methods provide camera poses of sufficient quality for the subsequent 3DGS reconstruction stages.

3.1 Problem

The 3DGS reconstruction pipeline requires each acquired image to be associated with a camera pose in a consistent world coordinate frame. These poses are needed both to align the image data during reconstruction and to relate the reconstructed object to the physical image acquisition setup.
The practical pose estimation problem is that the system uses two separate cameras. The tracking camera is used for ArUco-based pose estimation, while the capture camera is used for image acquisition. Therefore, estimating the pose of the tracking camera is not sufficient on its own. The pose must also be transformed into the corresponding capture camera pose in the world frame used throughout the workflow.
Pose estimation is treated as an enabling subproblem rather than the main contribution of the thesis. The purpose is therefore not to perform a broad benchmark of pose estimation methods, but to validate that the selected approaches provide pose information of sufficient quality for the PoC workflow. This is important because both 3DGS reconstruction and NBV planning depend on camera poses being expressed consistently in the same reference frame.
To address this, two pose estimation approaches are considered: ArUco-based pose estimation and SfM. ArUco-based pose estimation provides metric camera poses from known marker geometry, while SfM can estimate and refine camera poses from image correspondences. The chapter evaluates these methods, and a combined ArUco + SfM approach, under controlled conditions.

3.2 Theory

This section presents the theory needed to formulate and evaluate the pose estimation subproblem. First, the coordinate frames, pose transformations, and camera intrinsics are defined. Then, the pose estimation methods used in the project are described.

3.2.1 Camera model and coordinate frames

This section defines the camera projection model and the coordinate frames used throughout the project. Pose estimation and viewpoint planning both rely on transforming points and directions between frames, and a consistent notation is therefore important for clarity.

3.2.1.1 World frame definition

The global reference frame $\{W\}$ is defined by the ArUco marker with ID 0. All ArUco marker corners are defined as corners 0, 1, 2, and 3, which corresponds to the top-left, top-right, bottom-right, and bottom-left corners of each marker, respectively. The origin is located at the marker centre. The x_w-axis points parallel to the vector from corner 0 to corner 1, the y_w-axis points parallel to the vector from corner 3 to corner 0, and the z_w-axis is given by the right-hand rule, pointing outward from the marker plane. This choice provides a repeatable, physically meaningful, and easy-to-interpret world frame.

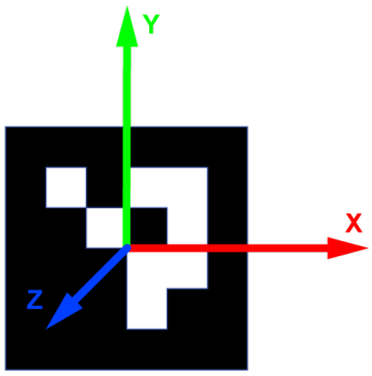


Figure 3.1: Definition of the world frame based on the reference ArUco marker.

3.2.1.2 Quadrotor and camera frames

Two cameras are mounted on the quadrotor, resulting in two camera frames. The tracking camera frame $\{C_t\}$ is used to detect ArUco markers and estimate pose relative to $\{W\}$. The capture camera frame $\{C_c\}$ is used for image acquisition for the 3DGS reconstruction.
The relationship between two frames $\{A\}$ and $\{B\}$ is described by the extrinsic transform matrix

$$T_{AB} = \begin{bmatrix} R_{AB} & t_{AB} \\ 0_{1\times 3} & 1 \end{bmatrix}, \tag{3.1}$$

where R_{AB} is a 3×3 rotation matrix and t_{AB} is a translation vector.
This transform maps a 3D point expressed in frame $\{B\}$ into frame $\{A\}$:

$$\begin{bmatrix} p_A \\ 1 \end{bmatrix} = T_{AB} \begin{bmatrix} p_B \\ 1 \end{bmatrix}. \tag{3.2}$$

Here, R_{AB} expresses the orientation of frame $\{B\}$ with respect to frame $\{A\}$, meaning that it rotates vectors from $\{B\}$ into $\{A\}$. The vector t_{AB} gives the position of the origin of frame $\{B\}$ expressed in frame $\{A\}$. To map from $\{A\}$ back to $\{B\}$, the inverse transform is used:

$$T_{BA} = T_{AB}^{-1}. \tag{3.3}$$

When using OpenCV's ArUco pose estimation, the returned rotation and translation describe the transform from the detected marker frame to the tracking-camera frame. Using the notation in this thesis, this transform is written as T_{C_tM}, where M is the detected marker frame. Since the world frame W is defined by the ArUco marker with ID 0, a detection of this marker gives

$$T_{C_tW}.$$

This transform maps points from the world frame into the tracking-camera frame. For the reconstruction and viewpoint-planning, the inverse relation is needed, since camera poses must be expressed in the world frame. The tracking-camera pose in the world frame is therefore obtained as

$$T_{WC_t} = T_{C_tW}^{-1}.$$

3.2.1.3 Pose composition

The image used for 3DGS reconstruction is acquired by the capture camera, while the ArUco-based pose estimate is obtained from the tracking camera. Therefore, the tracking-camera pose must be transformed into the corresponding capture-camera pose.

Let T_{WC_t} denote the pose of the tracking camera in the world frame, and let $T_{C_tC_c}$ denote the rigid transform from the capture-camera frame $\{C_c\}$ to the tracking-camera frame $\{C_t\}$. The capture-camera pose in the world frame is then computed by composing the two transforms.

$$T_{WC_c} = T_{WC_t} T_{C_tC_c}.$$

Since the capture camera is mounted on a servo, $T_{C_tC_c}$ is not a fixed matrix but depends on the current servo angle. The base transform at the neutral servo angle is derived from the CAD model of the platform. At any other servo angle, the transform is computed by applying the corresponding rotation around the servo pivot point.

This expression follows directly from the transform convention used in this thesis:$T_{C_tC_c}$ maps points from capture-camera frame to the tracking-camera frame, and T_{WC_t} maps points from the tracking-camera frame to the world frame. Their product therefore maps points from the capture-camera frame to the world frame.

In the remainder of the thesis, the term camera pose refers to the capture-camera pose T_{WC_c} when discussing image acquisition, COLMAP reconstruction, 3DGS training, and NBV planning, unless explicitly stated otherwise.

3.2.1.4 Camera intrinsics and calibration

In addition to the camera extrinsics, the camera intrinsic matrix is required. The intrinsic matrix K maps normalized image coordinates to pixel coordinates:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.4}$$

Here, f_x and f_y are the focal lengths in pixels, c_x and c_y define the principal point, and s is the skew.
Since the system uses two cameras, each camera has separate intrinsic parameters, denoted K_t and K_c. These parameters must be used consistently depending on which camera acquired the image. Camera calibration is performed using a chessboard pattern with OpenCV's implementation of Zhang's calibration method [8], which estimates the intrinsic parameters and lens distortion coefficients from multiple images of the patterns acquired at different orientations.

3.2.2 Pose estimation methods

This section presents the pose estimation methods used in the project. ArUco-based pose estimation is used to obtain metric camera poses from known marker geometry, while SfM is used to estimate camera poses and sparse 3D structure from image correspondences.

3.2.2.1 ArUco-based pose estimation

ArUco-based pose estimation is a model-based method for estimating a camera's 6D pose, i.e., its 3D position and 3D orientation, relative to a reference object with known geometry [4]. The method uses physically placed reference markers designed to be detected reliably and identified unambiguously. This differs from SfM, which infers pose from natural features in the scene.
An ArUco marker has a black and white square binary pattern with a black border, where the internal binary pattern encodes a unique ID. The detection pipeline begins by finding approximately square contours in the image and selecting candidate regions. Each candidate region is then warped into a canonical square image using a perspective transformation, and the binary pattern is decoded. If the decoded pattern matches an entry in the predefined marker dictionary, the marker is accepted and its ID is known.
Once a marker is detected, OpenCV provides the image coordinates of its four corner points. These 2D corner points correspond to known 3D corner points in the marker's own coordinate system. Since the marker is planar and its physical size is known, the marker corners provide metric 3D–2D correspondences.
With known correspondences between 3D marker corners and 2D image corners, pose estimation becomes a Perspective-n-Point (PnP) problem. The PnP problem estimates the rotation and translation that best explain how the 3D marker points project onto the 2D image, using the camera intrinsic matrix. Most commonly, the PnP result represents the transform from the marker coordinate system to the

camera coordinate system. If a world reference frame is defined using a specific marker, the camera pose in the world frame can then be obtained by chaining transforms between frames. In OpenCV, the PnP result is returned as a rotation vector and translation vector that transform 3D points from marker coordinate frame into the camera coordinate frame. In the notation used here, this corresponds to $T_{C_t M}$

A key advantage of ArUco-based pose estimation is that it provides metric scale directly, since the marker size is known. It is also computationally efficient and robust in controlled environments. However, its performance depends on marker visibility and image quality. Accuracy typically decreases with distance, and detection can fail under motion blur, strong reflections, or poor lighting.

3.2.2.2 Structure from Motion

SfM is a technique for estimating 3D structure and camera poses from a set of overlapping Two-Dimensional (2D) images [6]. The method identifies corresponding image features across multiple views and uses these correspondences to estimate the geometric relationship between images. Once the camera poses are known, corresponding image points can be triangulated to estimate their 3D positions, resulting in a sparse point cloud of the scene.

A common feature detector used in SfM pipelines is the Scale-Invariant Feature Transform (SIFT) [9]. SIFT detects keypoints that are designed to remain distinctive under changes in scale and rotation. It does this by constructing a scale space from Gaussian-blurred versions of the image,

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

where $I(x, y)$ is the image, $G(x, y, \sigma)$ is a Gaussian kernel, and σ is the scale. A Difference of Gaussians (DoG) representation is then computed as

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma),$$

where k is a constant scaling factor.

Local extrema in the DoG representation are used as candidate keypoints. Each pixel is compared to its neighbours in both image space and scale space, as illustrated in Figure 3.2. The selected keypoints are assigned dominant orientations and encoded using descriptor vectors, making it possible to match corresponding keypoints across multiple images.

The matched keypoints allow the relative camera poses to be estimated. After this, overlapping keypoints observed in multiple images can be triangulated to compute their 3D positions. The resulting sparse point cloud and camera poses are initially expressed in an arbitrary coordinate system, which must be aligned to the world frame if metric poses are required.

3.2.2.3 Georeferencing

Georeferencing for SfM is the process of transforming an arbitrary reconstruction coordinate system into a world coordinate system. The SfM reconstruction may

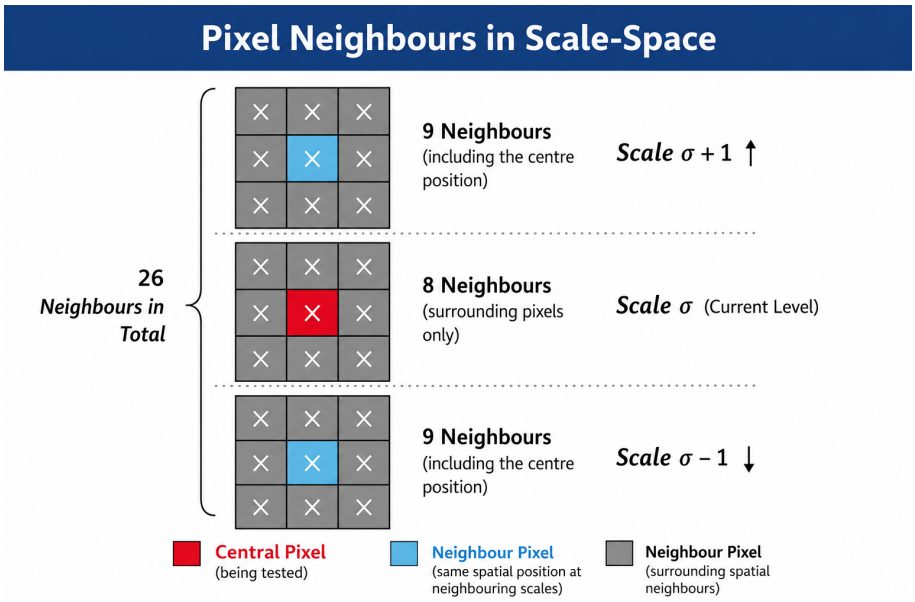


Figure 3.2: Pixel neighbours in scale-space. Created by the author with assistance from ChatGPT, OpenAI, on 4 June 2026.

have unknown scale, rotation, and translation, and these must be corrected before the reconstructed camera poses can be used together with the rest of the workflow. In this thesis, georeferencing is performed by using known camera poses from selected acquired images. These known poses define correspondences between camera positions in the local SfM reconstruction and the global world frame. A global similarity transform is then calculated and applied to the reconstructed camera poses and 3D points. The transform consists of a scale factor, a rotation matrix, and a translation vector, and it minimizes the error between corresponding camera positions in the local and global coordinate systems. After georeferencing, the sparse point cloud and reconstructed camera poses are expressed in the ArUco-defined world frame $\{W\}$. This makes the SfM reconstruction consistent with the physical image acquisition setup and allows the same coordinate frame to be used for 3DGS training and NBV planning.

3.3 Method

This section describes the experimental procedure used to assess whether the selected pose estimation approaches provide pose information of sufficient quality for the later 3DGS reconstruction pipeline. Three cases were considered: ArUco-based estimation, SfM-based estimation, and a combined ArUco + SfM approach.

3.3.1 Experimental setup

A rotational test rig was designed and constructed to validate the pose estimation approaches under controlled and repeatable conditions. The test rig provides known evaluation positions, which makes it possible to compare estimated poses against

ground-truth poses. The finished CAD model of the rig is shown in Figure 3.3.
The rig was developed in CAD and was inspired by both a 360° photo rig and the circular image acquisition motion of a quadrotor around an object. The structure was built using aluminium extrusions and 3D-printed components to provide rigidity, geometric repeatability, ease of assembly, and rapid manufacturing.
A rotating platform is mounted at the end of the radial arm and rotates around a fixed central platform where the object of interest is placed. The rig only allows rotation around the vertical axis while maintaining constant height. No pitch or roll rotation is permitted during operation. This makes the setup suitable for controlled evaluation of horizontal viewpoint changes similar to those considered in the thesis.
To ensure repeatable angular positions across experiments, the square aluminium extrusions were used as mechanical alignment references. A straight edge was used to position the rig at precisely 0°, 90°, 180°, and 270° for all validation measurements. These four positions were used as evaluation points in all experiments.
Ground-truth poses for the evaluation positions were defined from the known geometry of the rig in the CAD model together with the mechanically aligned angular positions. This provided a common reference against which the estimated poses were compared.

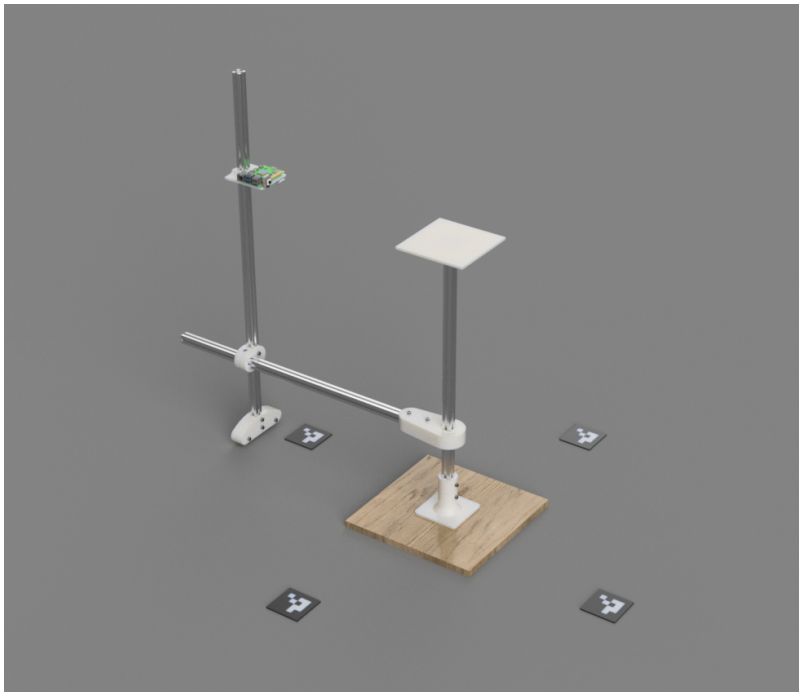


Figure 3.3: CAD model of the test rig.

Two cameras were rigidly mounted on the rotating platform together with a RPi. One downward-facing tracking camera was used for ArUco marker detection, and one capture camera was used for image acquisition in the SfM experiments. The base transform between the tracking camera and the capture camera is derived from the CAD model at the natural servo angle. During experiments the transform was updated according to the current servo angle. This matrix allows the pose estimated from the tracking camera to be transformed into the capture camera pose.

All experiments were carried out indoors under controlled conditions with stable lighting and a static background.

3.3.2 ArUco-based pose estimation experiment

To validate the ArUco-based pose estimation, the rig was positioned at the four evaluation angles 0°, 90°, 180°, and 270°. At each position, an image was acquired together with a pose estimate. The pose estimation was performed as described in Section 3.2.2.1. The experiment was repeated multiple times under stable lighting and static background conditions.

3.3.3 SfM-based pose estimation experiment

For the SfM-based validation, 36 images were acquired at evenly spaced angular increments of 10° over a full rotation.
To establish metric scale and align the reconstruction to the world coordinate system, the images at 0°, 90°, 180°, and 270° were assigned reference poses from the ground-truth defined by the rig geometry. These reference poses were used to align the SfM reconstruction, while the remaining images were reconstructed from image correspondences alone.
SfM then estimated poses for all 36 images, as well as the scaling and alignment of the coordinate system, using the four fixed poses. The estimated poses corresponding to the four evaluation positions were extracted and compared to the ground-truth values.

3.3.4 Combined ArUco + SfM experiment

A third experiment was conducted to evaluate the combined use of ArUco and SfM. The same 36 images were used as in the previous SfM experiment. However, instead of assigning the ground-truth poses to the four reference images, the poses estimated from the ArUco markers were used.
These four ArUco-derived poses were used in the SfM pipeline to provide scale and coordinate-frame alignment. The estimated poses corresponding to the four evaluation positions were extracted and compared to the ground-truth values.

3.4 Validation

This section presents the validation results from the pose estimation experiments described in Section 3.3. The validation compares three cases: ArUco-based pose estimation, SfM-based pose estimation, and a combined ArUco + SfM approach.
The estimated poses were compared with the ground-truth poses defined from the test-rig geometry at the four evaluation angles 0°, 90°, 180°, and 270°. Translational errors are reported independently for the x-, y-, and z-axes, while rotational error is reported only around the z-axis, since the rig constrains motion in the remaining rotational degrees of freedom. As a single summary measure, the mean Euclidean position error $\bar{e}$ is also reported for each method, defined as the mean over the four

evaluation angles of the Euclidean distance between the estimated and ground-truth camera positions. Unlike the per-axis errors, $\bar{e}$ describes the overall magnitude of the position error.

3.4.1 ArUco-based pose estimation

Table 3.1 summarizes the translation and rotation errors obtained using ArUco-based pose estimation.

Table 3.1: Translation and rotation errors for ArUco-based pose estimation.

Angle	e_x (mm)	e_y (mm)	e_z (mm)	e_ψ (deg)
0°	-20.0	76.4	6.3	1.4
90°	-1.0	-16.9	3.1	1.0
180°	-17.7	5.9	2.2	-5.2
270°	47.0	-21.1	0.9	-0.4

The mean absolute translation errors were $|e_x| = 21.43$ mm, $|e_y| = 30.08$ mm, and $|e_z| = 3.13$ mm. The mean Euclidean position error was $\bar{e} = 41.69$ mm. The mean absolute rotation error was $|e_\psi| = 2.0°$.
The results show that ArUco-based pose estimation provides metric pose estimates with relatively small height and yaw errors, but with larger variation in the horizontal position estimates. This indicates that the method is suitable as a practical pose source in the PoC system, although its translation accuracy is limited under the tested conditions.

3.4.2 SfM-based pose estimation

Table 3.2 summarizes the translation and rotation errors obtained using SfM-based pose estimation.

Table 3.2: Translation and rotation errors for SfM-based pose estimation.

Angle	e_x (mm)	e_y (mm)	e_z (mm)	e_ψ (deg)
0°	-4.2	-4.1	-1.5	2.0
90°	1.8	17.1	3.0	2.6
180°	-6.5	-11.1	-3.0	0.7
270°	-0.4	5.6	3.0	-5.4

The mean absolute translation errors were $|e_x| = 3.23$ mm, $|e_y| = 9.48$ mm, and $|e_z| = 2.63$ mm. The mean Euclidean position error was $\bar{e} = 10.77$ mm. The mean absolute rotation error was $|e_\psi| = 2.68°$.
Among the three evaluated approaches, the SfM-based method produced the lowest translational error. This indicates that, when reliable reference poses are available for alignment, SfM can provide highly accurate camera poses in the considered validation setup. The rotational accuracy is of the same order as for the ArUco-based method, although slightly worse in terms of mean absolute yaw error.

3.4.3 Combined ArUco + SfM-based pose estimation

Table 3.3 summarizes the translation and rotation errors obtained using the combined ArUco + SfM approach.

Table 3.3: Translation and rotation errors for combined ArUco + SfM-based pose estimation.

Angle	e_x (mm)	e_y (mm)	e_z (mm)	e_ψ (deg)
0°	-31.1	7.9	7.5	3.2
90°	3.1	41.6	14.5	3.9
180°	-42.3	-18.1	5.7	5.6
270°	-11.0	24.2	2.7	4.5

The mean absolute translation errors were $|e_x| = 21.88$ mm, $|e_y| = 22.95$ mm, and $|e_z| = 7.60$ mm.The mean Euclidean position error was $\bar{e} = 37.55$ mm. The mean absolute rotation error was $|e_\psi| = 4.3°$.
The combined approach did not improve the pose estimates compared with the SfM-based method using ground-truth reference poses. Compared with ArUco-based pose estimation, it reached a somewhat lower mean Euclidean position error, but this improvement was not consistent across all axes, and the rotation error was higher than for the ArUco-based method. This indicates that the accuracy of the reference poses used for alignment has a strong influence on the final SfM result, rather than that SfM reliably improves the ArUco-based pose estimates.
Overall, the validation shows that SfM with ground-truth reference poses gives the most accurate translation estimates in the controlled test-rig setup. However, ArUco-based pose estimation remains the practical pose source for the PoC workflow because it provides direct metric poses from onboard marker observations. The combined ArUco + SfM result shows that the accuracy of the reference poses used for georeferencing strongly affects the aligned SfM reconstruction. This limitation should therefore be considered when using pose estimation as input to the later 3DGS reconstruction and NBV planning stages.

4

3D Reconstruction

This chapter addresses the 3D reconstruction subproblem of the project. It describes how acquired images and associated capture-camera poses, expressed in the ArUco-defined world frame $\{W\}$, are transformed into a usable 3DGS model of the object. The chapter follows the pose estimation chapter and provides the 3DGS model used both as a reconstruction output and as input to the NBV planner.

4.1 Problem

The 3D reconstruction subproblem is to generate a usable 3DGS model from acquired images and their corresponding capture-camera poses T_{WC_c}. After COLMAP alignment, the sparse point cloud, camera poses, and resulting 3DGS model are represented in the ArUco-defined world frame $\{W\}$. In this thesis, the 3DGS model must represent the observed object with sufficient quality to support later evaluation, while also providing geometric information that can be used by the NBV planner.

This subproblem is challenging because the quality of the 3DGS reconstruction depends on several factors in the input data. The acquired images must have sufficient resolution, focus, lighting, and limited motion blur. They must also be acquired from viewpoints with enough overlap to support feature matching, sparse reconstruction, and camera pose alignment.

The reconstruction pipeline first generates a sparse point cloud, which provides an initial geometric representation of the scene and is used to initialize the 3DGS model. The resulting 3DGS model is then trained from the acquired images and camera poses. Since the NBV planner later uses the current 3DGS model to select new viewpoints, errors or missing regions in the reconstruction may affect the usefulness of the proposed viewpoints.

4.2 Theory of 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) is an explicit 3D scene representation and rendering method that represents a scene using a set of volumetric primitives called 3D Gaussians [7]. In contrast to implicit neural representations such as Neural Radiance Fields (NeRF) [10], which encode scene appearance in network weights, 3DGS uses an explicit set of parametrized primitives that support efficient rasterization-based rendering. The inputs to a 3DGS workflow are acquired images and corresponding

camera poses. In many workflows, a sparse reconstruction is first generated and used to initialize the 3DGS model.
Each Gaussian G_i is parametrized as

$$G_i = (\mu_i, \Sigma_i, c_i, \alpha_i), \tag{4.1}$$

where μ_i is the 3D mean position, Σ_i is the covariance matrix that determines the Gaussian's shape, size, and orientation, c_i is the colour, and α_i is the opacity.
The covariance matrix determines the spatial extent and orientation of each Gaussian. It is expressed as

$$\Sigma = RSS^\top R^\top, \tag{4.2}$$

where R is a rotation matrix and $S = \mathrm{diag}(s_x, s_y, s_z)$ is a scale matrix. The scale matrix determines the extent of the Gaussian along each axis. Initially, each Gaussian can be interpreted as a sphere, which corresponds to the case $S = \lambda I$.
Rendering is performed by projecting the Gaussians into the image plane and accumulating their colour and opacity contributions. Since each Gaussian has an opacity value $\alpha \in [0, 1]$, the final pixel colour depends on both overlap and visibility. The rendered pixel colour is computed using α-blending in a front-to-back accumulation scheme:

$$C = \sum_i T_i \alpha_i c_i, \tag{4.3}$$

$$T_i = \prod_{j<i} (1 - \alpha_j), \tag{4.4}$$

where C is the pixel colour and T_i is the transmittance up to Gaussian i.
To represent view-dependent appearance, 3DGS uses spherical harmonics (SH). A unit direction vector $\mathbf{v}$ represents the viewing direction, and the colour along that direction is approximated as

$$c(\mathbf{v}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{l,m} Y_{l,m}(\mathbf{v}), \tag{4.5}$$

where $Y_{l,m}$ are spherical harmonic basis functions and $c_{l,m}$ are learned coefficients. These coefficients allow the 3DGS model to represent view-dependent colour variations.
The 3DGS model is trained by comparing rendered images with the acquired images. The Gaussian parameters are refined through gradient-based optimization by minimizing a reconstruction loss. A simplified squared image loss can be written as

$$\mathcal{L} = \sum_{\text{pixels } p} \left| C_p - (I_{\text{obs}})_p \right|^2, \tag{4.6}$$

where $\mathcal{L}$ is the total loss, C_p is the rendered pixel colour, and $(I_{\text{obs}})_p$ is the observed colour in the acquired image.
The parameters are updated using gradient descent according to

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial \theta_i}, \tag{4.7}$$

where $\theta_i = (\mu_i, \Sigma_i, c_i, \alpha_i)$, η is the learning rate, and t is the iteration index.
Adaptive density control improves the representation during training. This process removes Gaussians that contribute little to the reconstruction and adds or subdivides Gaussians in regions where higher detail is needed. The optimization and density control steps are repeated until a stopping criterion is met, such as a fixed number of iterations or convergence. The final output is a trained 3DGS model that can be rendered from novel viewpoints.

4.3 Method

This section presents the reconstruction pipeline used in this thesis. The pipeline consists of two main stages: sparse point cloud generation and 3DGS training. The sparse point cloud stage provides an initial geometric representation and aligned camera poses, while the 3DGS training stage produces the trained 3DGS model used in later evaluation and NBV planning.

4.3.1 Sparse point cloud generation

The purpose of the sparse point cloud stage is to provide an initial geometric representation of the scene and camera pose alignment for 3DGS training. The input to this stage consists of the acquired images from the capture camera and the corresponding capture-camera poses T_{WC_c}. These poses are obtained by estimating the tracking-camera pose from ArUco marker observations and then applying the calibrated rigid transform between the tracking camera and the capture camera. The output is a sparse point cloud, camera parameters, and camera poses aligned to the ArUco-defined world frame $\{W\}$.
The sparse reconstruction is generated using the COLMAP SfM software. In this work, COLMAP is used as an external SfM tool, while the surrounding data preparation and processing steps are implemented within the proposed pipeline. COLMAP performs feature extraction, feature matching, and bundle adjustment to estimate camera poses and reconstruct a sparse set of 3D points representing the observed scene.
Before reconstruction, the images are undistorted using the known camera calibration parameters. This ensures that the intrinsic camera model used by COLMAP corresponds to the rectified images. The processed images are then used as input to the automatic reconstruction pipeline in COLMAP.
Since COLMAP first reconstructs the scene in a local SfM coordinate frame, the sparse reconstruction must be aligned to the ArUco-defined world frame $\{W\}$. The ArUco-based capture-camera poses are used as reference constraints for the COLMAP model aligner, which estimates a similarity transformation from the local SfM frame to $\{W\}$. To support this step and later processing, the pipeline includes routines for reading and writing the relevant COLMAP model files.
A region-of-interest (ROI) filtering step is applied after the COLMAP model has been aligned to the world frame. The purpose of this step is to reduce the influence of background structure and to define an object-centred region that can be reused by later reconstruction and viewpoint-selection stages. The ROI is defined as a cylinder

aligned with the world-frame Z-axis. Its horizontal centre is obtained from the centroid of the registered camera positions, since the cameras are acquired around the object. Its vertical coordinate is estimated from the sparse point cloud using a robust median-based estimate. This avoids placing the ROI too high when the camera trajectory is located above the object. The ROI radius is set as a fixed fraction of the minimum distance between the registered camera positions and their centroid, while the cylinder height is controlled separately by a height-scaling parameter.
In iterative runs, the ROI parameters are reused from the existing ROI file. This keeps the object boundary fixed when new NBV images are added, so that the ROI does not shift only because the camera set changes. Sparse points outside this cylindrical region are removed before the model is exported for 3DGS training.
The sparse reconstruction is then exported together with the corresponding camera parameters and poses. Background masking is also applied to prepare the images for 3DGS training. This is done because the aim of the reconstruction is to model the central object, while the surrounding background is not relevant for the final result. Masking is performed using rembg, a background-removal tool, using the BiRefNet-general segmentation model [11], which produces a binary foreground mask for each image. In the implemented pipeline, the computed ROI parameters are also saved and reused during later Gaussian post-processing and NBV candidate generation. The masked images, sparse point cloud, and corresponding camera poses are then used as input to the 3DGS training stage.

4.3.2 3DGS training

The purpose of the 3DGS training stage is to convert the sparse reconstruction into a denser and more visually informative 3DGS model of the target object. For this step, the open-source implementation provided by GraphDECO is used [7].
The inputs to the training stage are the acquired images, camera intrinsics, camera poses, and sparse point cloud generated in the previous stage. In this thesis, masked images are used so that the optimization focuses on the object of interest. Since the target object is located at the centre of the scene, the background is excluded from the training data.
Training is performed using the provided training script from the repository. The 3DGS model is trained for 30,000 iterations. Apart from minor configuration changes and the use of masked training images, the training procedure follows the original GraphDECO implementation.
At the start of training, the sparse point cloud and camera information are loaded and used to initialize the 3DGS model. The training then proceeds iteratively. At each iteration, one randomly selected camera view from the training set is selected, and the current 3DGS model is rendered from that viewpoint. The rendered image is then compared with the corresponding acquired image from the dataset. When masked images are used, the alpha mask is also applied during this comparison so that the optimization focuses on the target object rather than the removed background.
The optimization objective consists of an L_1 image loss and an SSIM-based term [12]. The L_1 loss measures the pixel-wise difference between the rendered and acquired

images, while SSIM evaluates similarity in overall image structure and local contrast. Using both terms makes the optimization sensitive to both exact pixel values and the overall visual consistency of the reconstruction.

During training, the learning rate is updated continuously, and the spherical harmonic degree of the Gaussian appearance model is increased every 1000 iterations. This allows the 3DGS model to start with a simpler appearance representation and gradually increase its ability to represent more detailed view-dependent colour variations.

The training script also includes densification and pruning. During optimization, the training procedure tracks which Gaussians are visible in each rendered view and how large they appear in image space. Based on this information, new Gaussians can be added in regions where the representation is too sparse, while Gaussians with limited contribution can be removed. The opacity values are also reset at predefined intervals in order to stabilize the optimization and avoid poor opacity distributions early in training.

After each iteration, the 3DGS model parameters are updated through gradient-based optimization. The final output of this stage is a trained 3DGS model of the target object.

4.3.3 3DGS model post-processing

After 3DGS training, the Gaussian model is post-processed to reduce background artifacts and keep the reconstruction focused on the target object. In the configuration used in this thesis, the post-processing consists of ROI-based pruning and mask-aware pruning.

The first post-processing step removes Gaussians whose centres lie outside the ROI defined during the sparse reconstruction stage. This applies the same object-centred spatial constraint to the trained 3DGS model as was used for the sparse COLMAP reconstruction. Since the ROI is cylindrical and aligned with the world-frame Z-axis, this step removes Gaussians that are located outside the expected object volume while preserving the central reconstruction region.

The second post-processing step uses the alpha masks from the masked training images. Each Gaussian centre is projected into the registered training cameras. For each view where the Gaussian is visible, the corresponding image mask is checked to determine whether the Gaussian projects into foreground or background. A Gaussian is removed if it is classified as background in a sufficiently large fraction of the views in which it is visible.

This mask-aware pruning step connects the 2D foreground segmentation used during training with the 3D Gaussian representation produced after training. It reduces residual background Gaussians that remain after image masking, while using observations from the registered camera views rather than a single image. The resulting model is therefore more object-centred before it is used for rendering, evaluation, and NBV planning.

4.4 Validation

This section verifies that the reconstruction pipeline produces usable intermediate outputs for the full thesis workflow. It does not evaluate the predefined and NBV-guided image acquisition strategies against each other; that comparison is performed in Chapter 7. Instead, the purpose here is to assess whether the sparse reconstruction and the trained 3DGS model are sufficiently coherent to support later NBV planning and evaluation.

4.4.1 Sparse point cloud

The sparse reconstruction provides a usable initialization for 3DGS training. Figure 4.1 shows the reconstructed sparse point cloud together with the estimated camera poses. The camera positions are arranged around the object in a consistent circular pattern, which indicates that the image set contains sufficient overlap and viewpoint variation for COLMAP to estimate scene structure and camera poses.
Although the reconstruction is sparse, it captures the overall shape and location of the object. This is sufficient for the intended purpose of this stage, since the sparse point cloud is primarily used to define the ROI and initialize the 3DGS reconstruction.
Overall, the results indicate that the sparse reconstruction step functions as intended and provides a suitable starting point for the following stage.

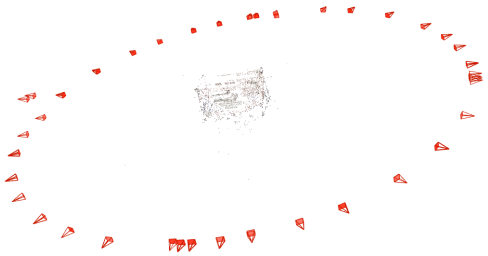


Figure 4.1: Sparse point cloud reconstruction with estimated camera poses.

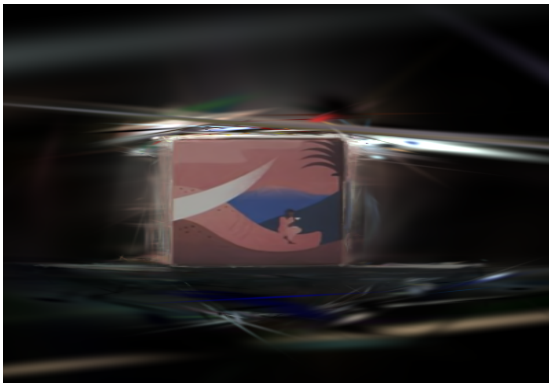
4.4.2 3DGS reconstruction

The trained 3DGS model provides a coherent visual representation of the target object. Figure 4.2 shows four views of the reconstructed object from different viewpoints. The 3DGS model captures the overall shape, proportions, and visual appearance of the object, which indicates that the reconstruction is not only valid from a single viewing direction.
The reconstruction quality is strongest for the lateral surfaces because all images in the dataset were acquired from the same height. Regions that require greater vertical viewpoint variation are less well represented. The limited image set likely also contributes to visible imperfections and distortions in the rendered results.

Despite these limitations, the 3DGS model is sufficiently detailed to represent the main surfaces and edges of the object. The main purpose at this stage is to verify that the 3DGS method can generate a usable representation from the available images.
Overall, the results indicate that the 3DGS training stage functions as intended and produces a 3DGS model that is suitable for later use in the full system.



(a) front view



(b) side view



(c) back view



(d) side view

Figure 4.2: 3DGS reconstruction results.

5
Next-Best-View Planning

This chapter addresses the NBV planning subproblem of the project. It describes how the current 3DGS model is used to evaluate candidate viewpoints and produce viewpoint recommendations for continued image acquisition. The NBV planner forms the adaptive part of the workflow, since future image acquisition is guided by the current state of the 3DGS reconstruction rather than by a fully predefined viewpoint sequence.

5.1 Problem

The NBV planning subproblem is to select new camera viewpoints that are expected to improve the current 3DGS model. In this thesis, the NBV planner should identify viewpoints that add useful information about poorly observed or uncertain regions of the object, while avoiding viewpoints that are too similar to already acquired views.
This is challenging because the usefulness of a viewpoint depends on the current reconstruction state. A viewpoint may be valuable if it observes uncertain object regions or surfaces that have not been sufficiently covered. However, a viewpoint may also be redundant if it observes regions that are already well represented in the 3DGS model. The NBV planner must therefore use the current reconstruction to estimate which candidate viewpoints are likely to improve the 3DGS reconstruction. The NBV planner is the mechanism that makes the image acquisition strategy adaptive. A predefined image acquisition strategy selects all viewpoints before the reconstruction is available, whereas the NBV planner selects new viewpoints based on the current 3DGS model. The purpose is therefore not only to select geometrically feasible viewpoints, but also to prioritize viewpoints that are expected to improve surface coverage and reconstruction quality.

5.2 Theory

This section presents the theory of NBV planning and the use of SfM and 3DGS in NBV planning.

5.2.1 Next-Best-View planning

Next-Best-View (NBV) planning is a strategy used in 3D reconstruction to determine the next camera viewpoint that is expected to provide the greatest improvement to

the current reconstruction [5]. The objective can be formulated in different ways depending on the reconstruction representation and the desired outcome.

Common NBV planning objectives include information gain, uncertainty reduction, surface coverage, visibility of unobserved regions, and reduced reconstruction or rendering error [5]. In this thesis, the objective is to select viewpoints that are expected to provide useful new information for the 3DGS reconstruction of a single object.

NBV planning can be formulated either as a discrete candidate-selection problem or as a continuous pose-optimization problem [5]. In a discrete formulation, a set of candidate viewpoints are generated, each candidate is evaluated using a utility function, and the highest-ranked candidate or candidates are selected. In a continuous formulation, the camera pose is treated as an optimization variable and updated directly based on the gradient of a differentiable objective function. This thesis uses a discrete candidate-based formulation because it is straightforward to constrain the candidate viewpoints to the physical image acquisition setup.

5.2.2 Reconstruction representations for NBV

The choice of reconstruction representation affects what the NBV planner can reason about. An SfM reconstruction consists of sparse 3D points and camera poses, which makes it computationally efficient and useful for estimating scene structure. However, a sparse point cloud does not explicitly represent continuous surfaces. This makes it difficult to determine whether a region is truly missing, poorly covered, or simply represented by few sparse points.

A 3DGS model provides a richer representation for NBV planning because it encodes 3D position, covariance, opacity, and appearance for a set of Gaussian primitives. This makes it possible to reason about which parts of the current 3DGS model have already been observed and which parts may still be uncertain. A 3DGS model can therefore provide a more informative basis for selecting new viewpoints than a sparse point cloud.

5.2.3 3DGS for NBV

Unlike SfM, 3DGS is geometry-aware, since it encodes 3D position, orientation, opacity, and density, which together provide a richer scene representation. This allows it to reason about uncertain or poorly observed parts of the current reconstruction. As a result, NBV planners based on 3DGS can more accurately target objectives such as uncertainty reduction and improved surface coverage.

For the purpose of guiding continued image acquisition in the 3DGS reconstruction workflow, relevant NBV planning strategies include multi-metric uncertainty, information gain, and gradient-based pose optimization. These strategies are useful when the goal is to improve metric accuracy, surface coverage, and visual completeness of poorly observed object regions.

The multi-metric uncertainty approach evaluates each candidate viewpoint using a utility function that combines several uncertainty-related terms. The utility of a

viewpoint q is denoted $U(q)$. As one illustrative example of such a formulation:

$$U_{\text{multi}}(q) = w_1 \cdot U_{\text{opacity}}(q) + w_2 \cdot U_{\text{coverage}}(q) + w_3 \cdot U_{\text{surface}}(q), \tag{5.1}$$

where $U_{\text{opacity}}(q)$ measures low-opacity regions visible from q, $U_{\text{coverage}}(q)$ measures newly visible surface areas, and $U_{\text{surface}}(q)$ measures surface uncertainty. The weights w_1, w_2, w_3 determine the relative importance of each term.
The information gain approach also uses uncertainty, but treats the 3D scene as a probabilistic model. In this setting, each part of the model is associated with a probability distribution that represents confidence in the scene parameters. The expected information gain of a viewpoint is then given by

$$U_{\text{MI}}(q) = H(\theta) - \mathbb{E}_{z_v}\Big[H(\theta \mid z_v)\Big], \tag{5.2}$$

where $U_{\text{MI}}(q)$ is the MI of viewpoint q, $H(\theta)$ is the current entropy of the scene model, θ is the probabilistic scene representation, and z_v is a hypothetical observation from the camera pose q. The second term is the expected remaining uncertainty after that observation.
The gradient-based approach optimizes the next camera pose continuously instead of selecting from a discrete set of candidates. If the utility function $U(q)$ is differentiable with respect to the camera pose q, the optimal pose can be obtained through gradient ascent:

$$q \leftarrow q + \eta \nabla_v U(q),$$

where η is the step size and $\nabla_v U(q)$ is the gradient of the utility function with respect to the pose. Out of these three approaches, the MI formulation (Equation 5.2) forms the basis of the NBV planner implemented in this thesis, as described in Section 5.3.1.

5.3 Method

This section describes how the NBV planner selects new candidate viewpoints from the current 3DGS model. The planner operates offline on a trained and post-processed 3DGS model produced by the reconstruction pipeline. It outputs a ranked set of candidate viewpoints that can be used for continued image acquisition.

5.3.1 NBV planner overview

The NBV planner takes the current 3DGS model, camera intrinsics, existing camera poses, and ROI information as input. The camera intrinsics define the projection used when candidate viewpoints are rendered and scored. The existing camera poses describe which viewpoints have already contributed to the current 3DGS reconstruction. The ROI defines the object centroid, which is used as the reference point for candidate viewpoint generation.
The implementation is based on GauSS-MI [3]. The MI scoring, Gaussian reliability representation, and modified rasterization interface follow the same general principle. However, the implementation used in this thesis is adapted to the reconstruction pipeline and physical image acquisition setup used in the project.

In contrast to the original GauSS-MI system, which is designed as an online active reconstruction framework, the implementation used here operates as an offline NBV planner. It uses an already trained 3DGS model, evaluates a discrete set of candidate viewpoints, and returns a ranked list of viewpoints for continued image acquisition. The execution of the recommended viewpoints is handled outside the planner.

5.3.2 Reliability update

The NBV planner first estimates which parts of the current 3DGS model are already well represented. The trained 3DGS model is loaded, and an initial reliability value is assigned to each Gaussian. The reliability is then estimated from the training images using a two-pass rendering procedure based on the GauSS-MI method [3].

For each training image, the scene is first rendered to produce an RGB output. The rendered image is compared to the corresponding ground-truth training image to compute a per-pixel reconstruction loss. This loss is transformed into a per-pixel quality signal, which is passed into a second render pass. The second pass accumulates a per-Gaussian reliability update as a log-odds value. Gaussians whose rendered appearance matches the training images well accumulate positive log-odds and are treated as reliable, while Gaussians that contribute to poorly-rendered regions accumulate negative log-odds and remain uncertain. The final per-Gaussian reliability is obtained by applying the sigmoid function to the accumulated log-odds.

This reliability update allows the NBV planner to distinguish between already well-observed regions and regions that may benefit from additional image acquisition. A candidate viewpoint receives a higher MI score when it observes uncertain Gaussians that are expected to add useful information to the 3DGS model.

5.3.3 Candidate viewpoint generation

After the reliability values have been updated, the NBV planner generates a set of candidate viewpoints around the object. The planner supports several candidate geometries; in the implemented workflow, the candidate viewpoints are constrained by the physical marker-circle setup used for image acquisition. This means that the candidate positions are generated around the marker-circle setup rather than freely over an unconstrained sphere.

The marker-circle formulation makes the candidate set consistent with the practical acquisition setup. It keeps the candidate viewpoints at a realistic distance from the object and within the region where image acquisition can be performed. The object centroid from the ROI is used as the main reference point when generating and orienting candidate viewpoints.

For each candidate position, the camera is oriented toward the object. The planner sweeps the vertical aiming point over the height range of the current Gaussian cloud, evaluating each pitch direction independently, and retains the orientation that produces the highest MI score.

5.3.4 Candidate evaluation and scoring

Each candidate viewpoint is evaluated using MI rendering. For each candidate, the 3DGS model is rendered with the modified GauSS-MI rasterizer, which uses the Gaussian reliability values to compute an information map [3]. Because the camera orientation is optimised per candidate by sweeping over multiple vertical aiming points, each candidate position is rendered several times and the orientation with the highest MI score is retained. The MI score is computed by summing the per-pixel information values over the rendered image. A high score indicates that the viewpoint is expected to observe uncertain regions of the 3DGS model.
The MI score is then adjusted using practical terms that reduce redundancy and encourage useful viewpoint diversity. First, a discount is applied to candidates that are close to existing camera poses and have a similar viewing direction. This prevents the NBV planner from selecting viewpoints that repeat already acquired observations. Second, candidates close to previously suggested NBV viewpoints can also be discounted, which helps avoid repeatedly proposing similar viewpoints in later planning iterations.
The final score of a candidate viewpoint q can be summarized as

$$S(q) = U_{\mathrm{MI}}(q)\,(1 - d_{\mathrm{existing}}(q))\,(1 - d_{\mathrm{tried}}(q))\,, \tag{5.3}$$

where $U_{\mathrm{MI}}(q)$ is the MI score, $d_{\mathrm{existing}}(q)$ is the discount for similarity to existing camera poses, $d_{\mathrm{tried}}(q)$ is the discount for similarity to previously suggested NBV poses.

5.3.5 Candidate ranking and output

After all candidates have been evaluated, they are ranked according to their final score. The planner returns a ranked set of candidate viewpoints rather than only a single viewpoint. This makes it possible to use several high-scoring and spatially diverse viewpoints for continued image acquisition.
The final output of the NBV planner consists of the ranked candidate viewpoints in the same world coordinate system as the COLMAP reconstruction. These viewpoints can then be used as recommendations for continued image acquisition.

5.4 Validation

This section verifies that the NBV planner can generate candidate viewpoints and select plausible viewpoints from a trained 3DGS model. It does not evaluate whether the NBV-guided image acquisition strategy outperforms the predefined strategy; that comparison is performed in Chapter 7. The purpose here is only to assess whether the planner produces reasonable viewpoint recommendations from the current 3DGS reconstruction.
The initial result of the NBV planning stage is shown in Figure 5.1. The existing camera poses, generated candidate viewpoints, and selected NBV viewpoint are visualized around the object. The figure shows an early planning result where the existing images were acquired mainly from lateral viewpoints.

Since the current image set mainly observes the object from the side, the upper region of the object is less well represented. The NBV planner responds to this by proposing viewpoints above the object, where additional observations are expected to provide useful new information. This indicates that the planner reacts plausibly to missing vertical coverage in the current 3DGS model.
Although this result is only an initial verification, it supports the feasibility of using the NBV planner to recommend new viewpoints based on the current 3DGS reconstruction. The result should therefore be interpreted as a qualitative check of the planning method, not as final evidence that the NBV-guided image acquisition strategy improves reconstruction performance.

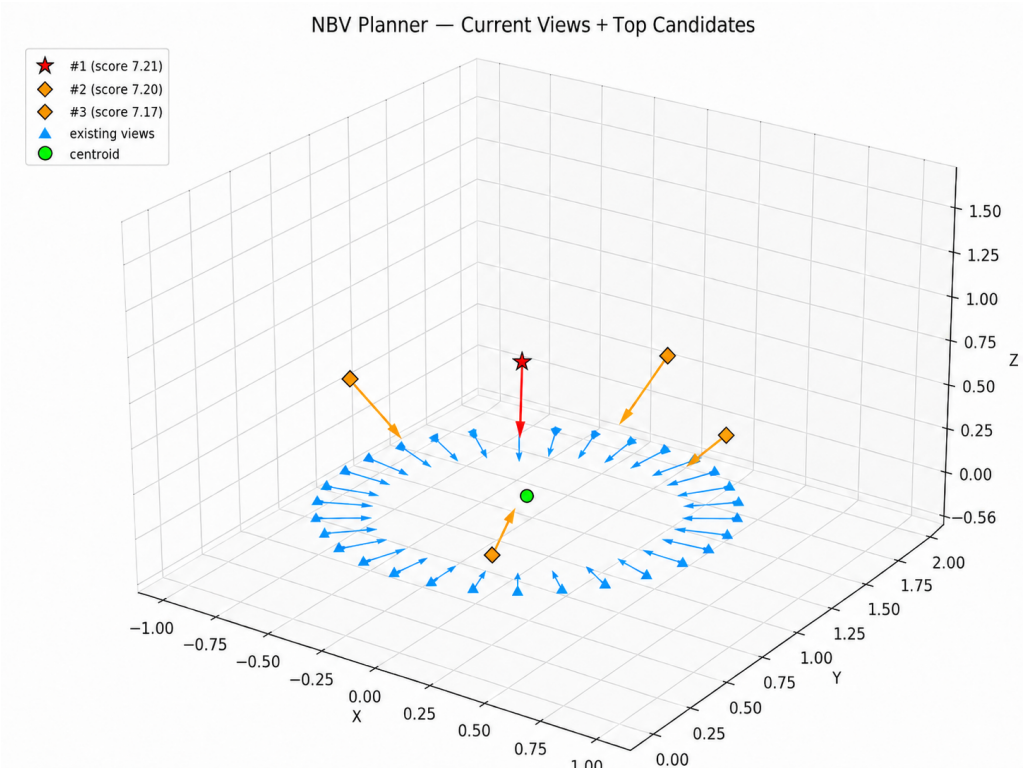


Figure 5.1: Visualization of the initial NBV planning result.

6
System Architecture

This chapter describes the implemented quadrotor-based PoC system used to connect the three subproblems addressed in this thesis: pose estimation, 3DGS reconstruction, and NBV planning. The system is based on a semi-autonomous workflow in which the quadrotor performs flight execution, pose estimation, and image acquisition, while 3DGS training and NBV planning are performed off-board. This division reflects both the hardware constraint that 3DGS training exceeds the the RPi's onboard capability and the deliberate PoC scope of evaluating the acquisition workflow rather than onboard real-time processing.
The purpose of the quadrotor system was to investigate how the different parts of an NBV-guided image acquisition workflow could be integrated on a physical platform. However, the implemented quadrotor system was not sufficiently reliable for repeatable autonomous indoor acquisition during the final quantitative comparison between predefined and NBV-guided image acquisition strategies. Therefore, the quadrotor is treated as a PoC platform, while the final acquisition strategy comparison is performed using the controlled proxy setup described in Chapter 7.
The chapter presents the hardware setup, software architecture, intended flight procedure, implementation status, and practical limitations of the quadrotor PoC.

6.1 Hardware setup

The hardware setup is built around the PX4 Development Kit X500 v2 from Holybro, shown in Figure 6.1. The platform provides the mechanical frame, propulsion system, power distribution, and flight-control hardware required for quadrotor-based image acquisition. The kit includes a Pixhawk 6C flight controller, M8N GPS, plug-and-play telemetry radio, four electronic speed controllers (ESCs), a power distribution board (PDB), a frame kit, and propellers. The Pixhawk 6C acts as the main flight controller and is responsible for stabilizing and controlling the quadrotor during flight. The Pixhawk connection ports are shown in Figure 6.2.
The onboard computing and image acquisition tasks are handled by an RPi 5 mounted on the quadrotor. The RPi communicates with the Pixhawk through UART, using the RPi GPIO pins and the Pixhawk's second telemetry port (TELEM2). This connection allows the RPi to exchange flight-related messages with the Pixhawk while also controlling the camera system used for pose estimation and image acquisition.
The image acquisition system consists of two cameras connected to the RPi. One camera is used for ArUco-based pose estimation, while the other is used for image



Figure 6.1: PX4 Development Kit X500 v2 from Holybro.

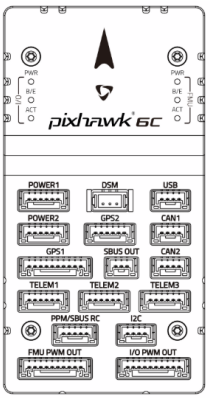


Figure 6.2: Pixhawk 6C.

acquisition for the 3DGS reconstruction. The capture camera is mounted on a servo-controlled frame, allowing the viewing direction to be adjusted during image acquisition. The detailed camera setup is described in Section 6.1.1. The ESCs receive their PWM signals from a PWM module connected to the Pixhawk's I/O PWM main output.
The communication between the quadrotor and the external operator computer is handled through the telemetry radio connected to TELEM1 on the Pixhawk. The telemetry radio communicates with a receiver connected to a laptop running QGroundControl (QGC). This setup allows missions to be uploaded and started, while also providing real-time monitoring of the quadrotor during operation.
The entire setup is powered by a 4S LiPo battery with a capacity of 6100 mAh and a C rating of 100C. The battery connects to the PDB and a power module, which powers the Pixhawk. The PDB powers the four ESCs and motors and also indirectly powers the RPi and micro servomotor through a DC-DC step-down converter with an output of 5 V and a maximum current of 5 A. The complete hardware flowchart is shown in Figure 6.3.

6.1.1 Camera setup

The camera setup consists of one tracking camera and one capture camera, each with a separate role in the image acquisition workflow. The tracking camera is used for ArUco marker detection and pose estimation, while the capture camera is used to acquire images for the 3DGS reconstruction.
Both cameras are 16 MP Arducam cameras connected to the RPi through the Camera Serial Interface (CSI). The cameras are independently calibrated because they are used for different tasks and have different optical requirements. The tracking camera uses a wider field of view (FOV) to increase marker visibility during flight. It is also configured at a lower resolution to reduce computational load, since high image resolution is not required for ArUco marker detection.
The tracking camera is mounted as a downward-facing camera on the quadrotor. This orientation allows it to detect ArUco markers placed on the ground and estimate the quadrotor pose relative to the defined world frame. The capture camera is

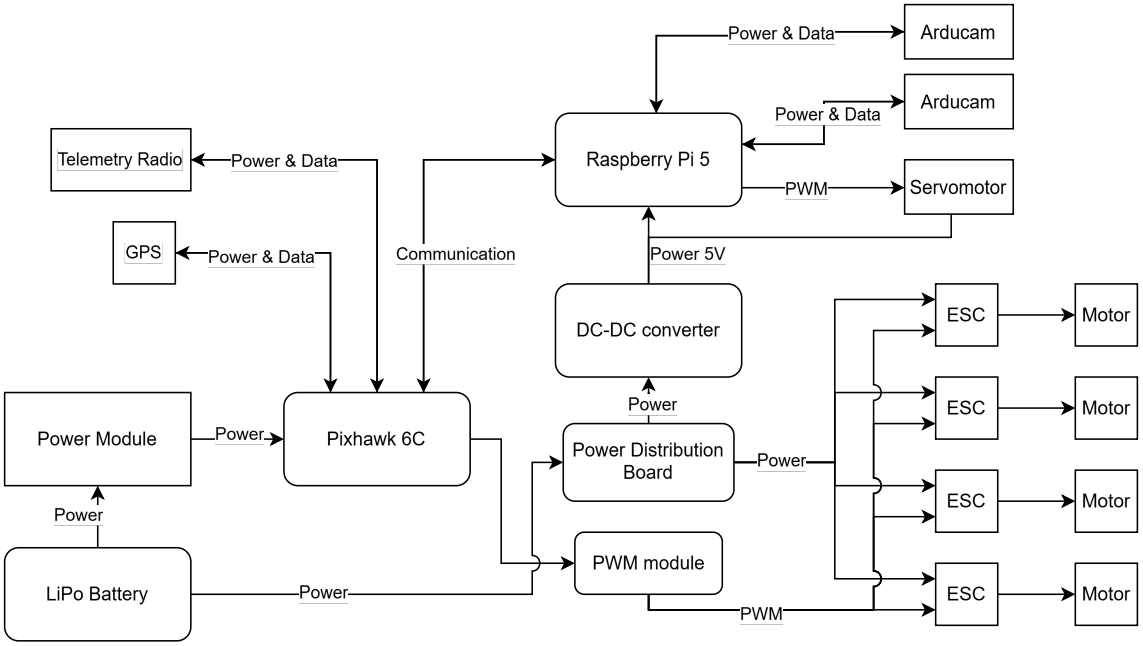


Figure 6.3: Flowchart of the quadrotor hardware.

mounted on an extended frame controlled by a servomotor, which allows images to be acquired from different pitch angles. This makes it possible to adjust the viewing direction of the capture camera without changing the quadrotor pitch.

6.2 Software architecture

The software architecture is semi-autonomous because the onboard system does not perform the full 3DGS reconstruction and NBV planning loop. During flight, the RPi and Pixhawk handle onboard image acquisition, pose estimation, and flight execution. The acquired images and poses are then transferred to an off-board GPU provided by Wirevision, where the 3DGS training is performed. The resulting 3DGS model is used as input to the configured GauSS-MI NBV planner, which outputs proposed viewpoints for continued image acquisition.
In the intended autonomous workflow, these viewpoints would then be provided to the RPi on the quadrotor, after which the quadrotor would acquire the next set of images. In the implemented PoC, this full acquisition–reconstruction–planning loop was designed and partially tested, but it was not sufficiently reliable for the final quantitative evaluation.
During flight, the onboard software system is distributed between the RPi and the Pixhawk, while an external computer running QGC is used as the ground-control interface. The RPi runs Raspberry Pi OS as its host operating system, which enables the use of the Arducam cameras. To run ROS 2 Jazzy Jalisco [13], an Ubuntu containerized environment is created on the RPi using Docker. The software is primarily written in C++ and uses a ROS 2 node-based architecture.
The RPi communicates with the Pixhawk over UART using transmit (TX) and receive (RX) lines. The Pixhawk runs PX4 firmware [14] for flight-control processing

and uses uORB for internal messaging. To exchange messages between the Pixhawk and the RPi, a DDS-based bridge is used. The PX4 firmware uses micro XRCE-DDS middleware to allow uORB messages to be published and subscribed on a companion computer in a way that is similar to ROS 2 topics. The micro XRCE-DDS client on the Pixhawk and the micro XRCE-DDS agent on the RPi provide the communication bridge between the PX4 and ROS 2 environments. The messages are therefore accessible within the Docker container on the RPi, enabling bidirectional communication between the RPi and Pixhawk, as shown in Figure 6.4.
QGC is used to operate and manage the quadrotor through telemetry communication with the Pixhawk. It provides real-time flight monitoring by displaying quadrotor position, altitude, speed, battery status, and other flight information. It is also used as the control interface where missions are uploaded and started. In addition, QGC provides access to Pixhawk configuration tools, including sensor calibration, parameter configuration, firmware management, and log analysis when needed.

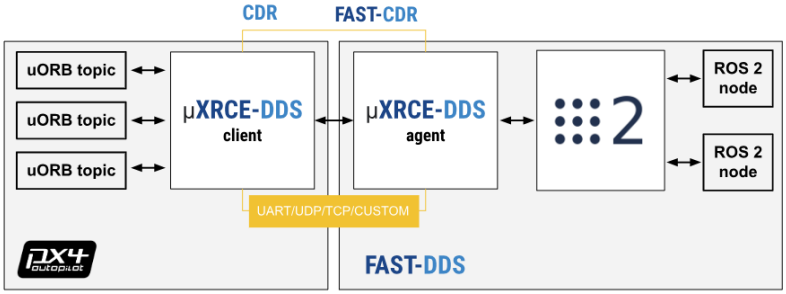
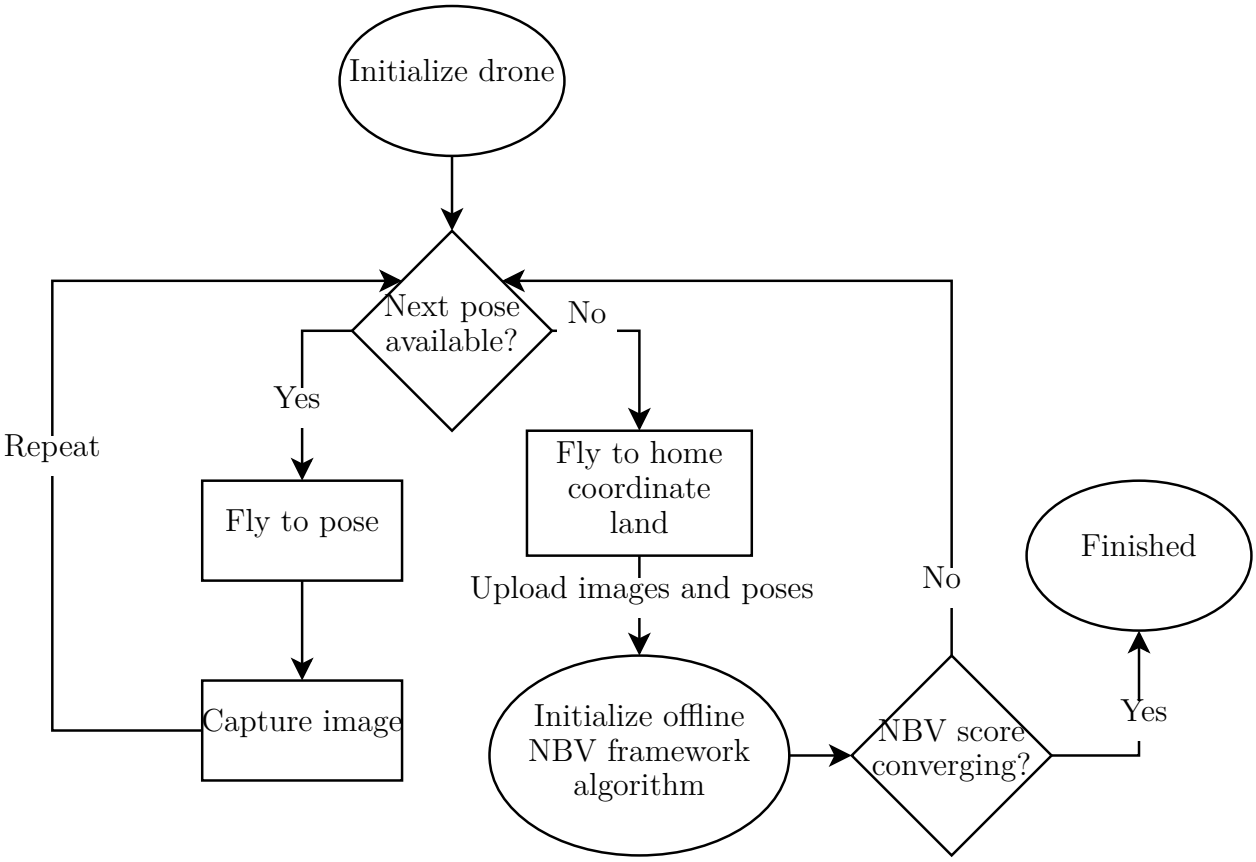


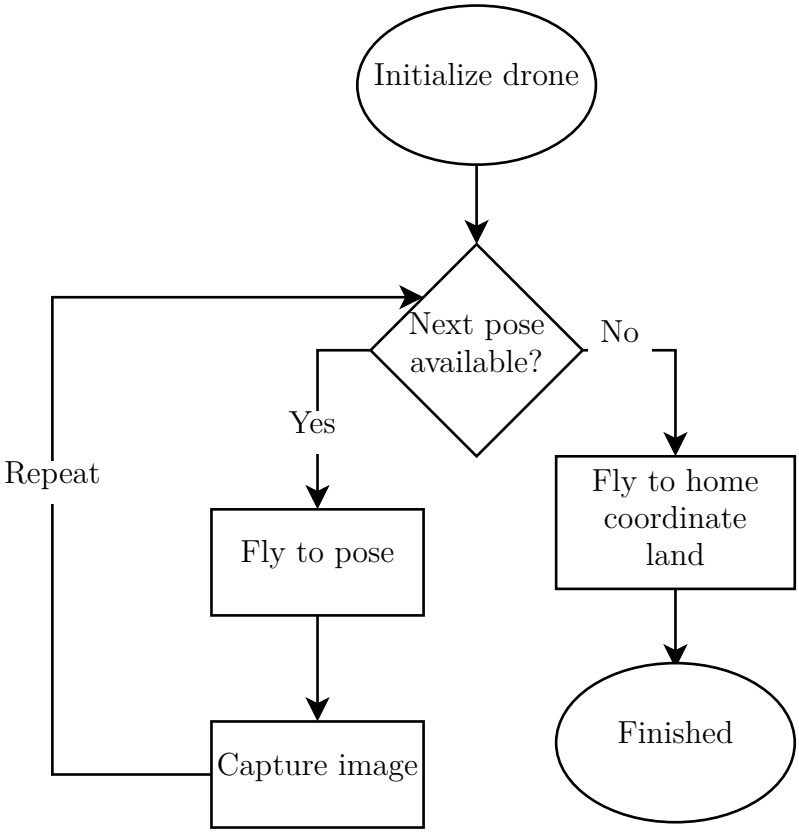
Figure 6.4: Client-agent communication between the Pixhawk and RPi.

6.3 Flight procedure

Two image acquisition procedures are considered in this thesis: a predefined procedure and an NBV procedure. In the predefined procedure, all camera viewpoints are selected before the flight. In the NBV procedure, an initial set of viewpoints is predefined, while the remaining viewpoints are selected by the NBV planner based on the current 3DGS model. These procedures are referred to as the predefined and NBV image acquisition procedures.
The procedures described in this section represent the intended quadrotor-based acquisition workflows. The final comparison between the predefined and NBV strategies is performed using the controlled proxy setup described in Chapter 7, while the quadrotor implementation is used to evaluate the feasibility and limitations of autonomous acquisition.
The flight procedures for both approaches are shown in Figure 6.5: the NBV-based procedure in Figure 6.5a and the predefined procedure in Figure 6.5b.
Both image acquisition procedures use the same custom flight-mode framework. To fly a custom mission, a custom flight mode is created and uploaded to the Pixhawk. The mode can then be enabled by the user in QGC. The control software uses two state machines: one for the internal custom flight states and one for handling QGC's



(a) NBV-based flight procedure.



(b) Predefined flight procedure.

Figure 6.5: Flowcharts of the two flight procedures.

built-in flight modes. The second state machine is required to switch between flight modes, since the quadrotor uses the built-in TakeOff and Landing modes in QGC. This allows the quadrotor to take off to a desired altitude, execute the custom flight mode, and land in a standardized way. The second state machine acts as a simple state executor and is not described further.
The custom flight state machine starts after the quadrotor has taken off and reached a hovering position. Its first objective is to detect the ArUco marker with ID 0 and center the quadrotor above it. Since the Pixhawk's position estimate is based on internal sensors and cannot be directly overwritten by the custom controller, the centering is performed by sending corrected waypoints to the quadrotor. The correction term is defined in the world coordinate frame and is based on the difference between the detected and expected position of the ArUco marker relative to the quadrotor. The objective is to make the relative error between the estimated and desired quadrotor position converge toward zero.
The centering state remains active until the quadrotor has reached its waypoint and held its position within a specified positional and angular error for a set time. When these criteria are met, the capture camera is triggered to perform image acquisition. After the first image has been acquired, the state machine enters the Fly Circle state, where the remaining predetermined viewpoints are visited and an image is acquired at each waypoint. When all waypoints have been reached and the images have been acquired, the second state machine calls the Landing mode, and the quadrotor lands at its starting position. This structure is used for the predefined image acquisition procedure and for the initial stage of the NBV image acquisition procedure.
In the NBV image acquisition procedure, the initial images and poses are used to create an initial 3DGS reconstruction. The 3DGS reconstruction follows the method described in Section 4.3. The current 3DGS model is then used as input to the NBV planner described in Section 5.3. The NBV planner outputs new viewpoints that are expected to improve the 3DGS reconstruction. In the intended autonomous workflow, these viewpoints would then be provided to the RPi, and the quadrotor would perform another image acquisition step from the proposed viewpoints. This acquisition-reconstruction-planning loop is intended to repeat untill the NBV scores converge, indicating that additional viewpoints are unlikely to improve the reconstruction further. In the current PoC implementation, convergence is not evaluated automatically; instead, the loop runs for a fixed number of stages, as described in the evaluation setup in Chapter 7.1

6.4 Implementation status

The implemented quadrotor PoC was evaluated at a system level to determine which parts of the intended workflow were implemented, which parts were partially verified, and which parts were not reliable enough for the final evaluation. This section does not present the final reconstruction results; those are presented in Chapter 7. Instead, it documents the practical outcome of the quadrotor implementation.
Table 6.1 summarizes the implementation status of the main system components.
The implementation showed that the main hardware and software components could be integrated into a common PoC architecture. The RPi could communicate with

Table 6.1: Implementation status of the quadrotor PoC system.

Component	Status	Verification
Quadrotor hardware platform	Implemented	Assembled with Pixhawk, RPi, cameras, telemetry, and power distribution.
RPi—Pixhawk communication	Implemented	UART and micro XRCE-DDS communication tested between PX4 and ROS 2.
Tracking camera and ArUco pose estimation	Implemented	Marker detection and pose estimation tested under controlled conditions.
Capture camera system	Implemented	Image capture and servo-mounted camera configuration tested.
Predefined quadrotor acquisition	Partially verified	Custom flight mode and waypoint-based acquisition logic implemented and tested.
NBV quadrotor loop	Partially verified	Off-board reconstruction and NBV planning integrated, but the full autonomous loop was not repeatable indoors.
COLMAP, 3DGS, and NBV pipeline	Implemented	Used for off-board reconstruction and NBV planning in the final strategy evaluation.

the Pixhawk, the cameras could be connected and controlled, ArUco-based pose estimation could be performed under controlled conditions, and the off-board reconstruction and NBV-planning pipeline could be executed. However, the complete autonomous workflow was not sufficiently robust to support a fair quantitative comparison between predefined and NBV image acquisition strategies.

6.5 Summary

The implemented quadrotor system demonstrates that the main hardware and software components required for NBV-guided image acquisition can be integrated into a common PoC architecture. The platform supported Pixhawk–RPi communication, camera-based image acquisition, ArUco-based pose estimation, and off-board reconstruction and NBV planning.
At the same time, testing showed that the system was not sufficiently reliable for repeatable autonomous indoor image acquisition. The limitations primarily affected indoor positioning and the integration of the full acquisition-reconstruction-planning loop. The practical limitations that prevented reliable autonomous operation are presented in Section 7.6.

7
Evaluation and Results

This chapter presents the results of both evaluations conducted in this thesis: the comparison between the predefined and NBV-guided image acquisition strategies, addressing RQ1 and RQ2, and then the practical limitations of the quadrotor PoC system, addressing RQ3. The purpose of the evaluation is to determine whether the NBV strategy improves 3DGS reconstruction quality compared with a predefined strategy.

To avoid ambiguity, this thesis distinguishes between the image acquisition strategies and the reconstruction modes. The acquisition strategy is either predefined where all viewpoints are selected before acquisition, or NBV-guided where all viewpoints are selected adaptively based on the current 3DGS reconstruction. The reconstruction mode is either stage-wise or batched.

The main comparison is performed at equal numbers of captured images in order to evaluate reconstruction quality as a function of acquisition progress. The predefined and NBV strategies are therefore compared after 12, 18, 24, 30, and 36 training images. This stage-wise comparison shows whether the NBV strategy reaches a given reconstruction quality with fewer images, or produces better reconstruction quality for the same number of acquired images.

In addition, the batch reconstructions are compared after both strategies have reached 36 training images. This batch comparison evaluates whether the NBV strategy provides a better final reconstruction when both methods use the same total image budget. The 36-image staged reconstruction is the last point in the staged comparison, while the batch reconstruction is trained from scratch using all 36 images.

The evaluation focuses on the acquisition strategy rather than on the autonomous flight performance of the quadrotor. As described in Chapter 6, the implemented quadrotor system was not sufficiently robust to complete the full acquisition and reconstruction workflow autonomously in a repeatable indoor setting. Therefore, the final comparison between the predefined and NBV strategies is performed using a controlled proxy setup. This makes it possible to compare the two strategies under repeatable conditions while still preserving the intended reconstruction and NBV planning workflow.

7.1 Evaluation objective

The objective of the evaluation is to assess whether the NBV image acquisition strategy can improve object-centred 3DGS reconstruction compared with a predefined

image acquisition strategy. The predefined strategy is used as a baseline because it represents a simple and repeatable acquisition method where all viewpoints are selected before reconstruction. The NBV strategy instead starts from an initial set of images and uses the current 3DGS model to select additional viewpoints.
The comparison is designed to evaluate whether the additional complexity introduced by NBV planning is justified by improvements in reconstruction quality, acquisition efficiency, or practical reconstruction performance. The evaluation is structured around RQ1 and RQ2 from Chapter 1. RQ3 concerns the practical limitations of the quadrotor PoC system and is addressed through the results in Section 7.6 and the discussion in Section 8.2.
The evaluation uses image registration results, image-based reconstruction quality metrics, reconstruction density indicators, practical reconstruction cost, and visual inspection.
Since no ground-truth geometry is available, the evaluation does not measure absolute geometric accuracy. Instead, visual completeness is assessed indirectly through rendered-view metrics, reconstruction density indicators, including the number of sparse COLMAP points and optimized Gaussians, and qualitative inspection of rendered views.

7.2 Evaluation setup

The final evaluation is performed in a controlled indoor environment using the controlled proxy setup. The proxy setup is a handheld platform carrying one tracking camera and one capture camera, both actuated by an onboard RPi when a capture button is pressed. It is designed to mimic the image acquisition of the quadrotor while allowing repeatable manual capture. The camera-to-camera transform was handled using the same approach: the base transform was derived from the CAD model at the neutral servo angle, and updated according to the current servo angle during acquisition. The tracking camera is used to estimate the camera pose, while the capture camera is used to acquire the images used for reconstruction. The capture camera on this platform is a separate 12 MP camera, not the 16 MP Arducam capture camera used on the quadrotor. Unlike the quadrotor cameras, the intrinsic parameters of this camera were not determined through a separate calibration procedure. Instead, COLMAP estimated the camera intrinsics directly from the acquired images during sparse reconstruction. During acquisition, the estimated pose of the capture camera was monitored. When the tracking camera could not detect a marker, the pose was instead estimated from the known acquisition geometry, and later retrieved through georeferencing.
Two objects are used in the evaluation. The first object is an office chair with a box placed on it and the second is the quadrotor used as a static reconstruction object (referred to as the quadrotor object in the following). For each object, two datasets are acquired: one using the predefined strategy and one using the NBV-guided strategy.
The predefined strategy consists of 36 training images acquired from viewpoints evenly spaced around the object in a circular motion. The viewpoints are distributed over three height levels, with 12 images captured at each height. This gives a fixed

and repeatable baseline dataset. For the stage-wise evaluation, predefined subsets are formed from the same fixed sequence at 12, 18, 24, 30, and 36 images in order to compare it to the staged NBV reconstructions.
The NBV strategy starts with 12 initial training images, from the second height level of the predefined strategy. An initial 3DGS model is trained from these images and used by the NBV planner to select six additional viewpoints. A new reconstruction is then trained upon the previous one using the updated dataset, after which the NBV planner selects another six viewpoints. This process is repeated until 24 additional images have been acquired, resulting in a total of 36 training images. The NBV strategy is therefore evaluated after 12, 18, 24, 30, and 36 images.
The predefined and NBV stage-wise strategies are compared at equal numbers of training images. This makes it possible to compare and evaluate whether the NBV-selected viewpoints improve reconstruction quality during acquisition. The batch reconstructions are also compared to evaluate the result when both strategies start from scratch with the collected images.
Five evaluation views are acquired for each object. These views are not included in any of the training datasets. They are used only for evaluating the trained 3DGS models. This makes it possible to compare how well the reconstructions generalize to viewpoints that were not used during training.
To ensure a fair comparison, the same object, camera setup, image resolution, lighting conditions, camera calibration, pose-estimation procedure, COLMAP pipeline, and 3DGS training settings are used for both acquisition strategies whenever possible. Reconstruction settings are kept fixed between the strategies and are summarized in Table 7.3.
The evaluated acquisition cases are summarized in Table 7.1, and the image counts and cumulative iterations used for the stage-wise comparison are summarized in Table 7.2.

Table 7.1: Evaluation cases for the predefined and NBV-guided image acquisition strategies.

Object	Strategy	Initial	Added	Total	Eval. views
Office chair with box	Batch Predefined	36	0	36	5
	Batch NBV	36	0	36	5
	Stage-wise Predefined	12	24	36	5
	Stage-wise NBV	12	24	36	5
Quadrotor	Batch Predefined	36	0	36	5
	Batch NBV	36	0	36	5
	Stage-wise Predefined	12	24	36	5
	Stage-wise NBV	12	24	36	5

7.3 Evaluation metrics

The evaluation metrics are chosen to answer the research questions without relying on measurements that cannot be obtained reliably in the final setup. Since no ground-truth geometry is available, geometric accuracy is not evaluated directly.

Table 7.2: Training-image counts and cumulative training iterations used in the staged reconstruction comparison.

Stage-wise Strategy	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
Predefined images	12	18	24	30	36
NBV images	12	18	24	30	36
Cumulative iterations	30k	40k	50k	60k	70k

Table 7.3: Fixed reconstruction and evaluation settings used for both strategies.

Setting	**Value**
Acquisition setup	Controlled proxy setup
Tracking camera	16 MP Arducam
Capture camera	12 MP camera (not the quadrotor Arducam)
Image resolution	1514 x 2022
Object mask/background handling	Rembg-based object masking; black background for both rendered and reference images at evaluation
COLMAP settings	Same settings used for both strategies
3DGS training iterations	30,000 for batch and initial reconstruction. 10,000 for stage-wise reconstructions
3DGS training settings	Same settings used for both strategies
Evaluation views	5 views not used during training
Training hardware	AMD Ryzen 7 5700X (8-core, 4.67 GHz boost); NVIDIA GeForce RTX 2080 Ti (11 GiB VRAM); 31 GiB RAM; Ubuntu 24.04.4 LTS

Metrics such as absolute surface error, ground-truth coverage, and geometric completeness are therefore excluded. Instead, the evaluation focuses on metrics that can be measured consistently for both acquisition strategies.

7.3.1 Image registration and reconstruction information

Before evaluating the trained 3DGS models, the number of successfully registered images is recorded. This is important because two strategies may acquire the same number of input images, while COLMAP may only register a subset of them. A lower number of registered images means that less of the acquired data contributes to the final reconstruction.

In addition to image registration, the number of sparse COLMAP points and the number of optimized 3D Gaussians are recorded. These values do not directly measure reconstruction quality, but they provide useful information about the density and size of the reconstruction. A larger number of sparse points or Gaussians may indicate that more visual structure was reconstructed, but it may also increase storage cost and does not necessarily guarantee better novel-view rendering quality.

7.3.2 Image-based reconstruction quality

Image-based reconstruction quality is evaluated by comparing rendered images from the trained 3DGS models with corresponding reference images. The reference images are acquired from five evaluation viewpoints that are not included in the training datasets. This makes it possible to evaluate how well each reconstruction generalizes to viewpoints outside the training set.

For each acquisition strategy, a separate 3DGS model is trained using the same reconstruction pipeline and training settings. The trained models are then rendered from the same evaluation viewpoints. The rendered images are compared with the corresponding reference images using Peak Signal-to-Noise Ratio (PSNR) [15], Structural Similarity Index Measure (SSIM) [12], and Learned Perceptual Image Patch Similarity (LPIPS) [16]. These metrics are used in the original 3DGS paper to evaluate novel-view synthesis quality [7].

PSNR measures pixel-wise similarity between the rendered and reference images. A higher PSNR value indicates a lower pixel-wise error. SSIM measures structural similarity by considering luminance, contrast, and local image structure. A higher SSIM value indicates better structural similarity. LPIPS measures perceptual similarity using learned visual features, where a lower value indicates better perceptual similarity.

These three metrics are used together because they capture different aspects of visual reconstruction quality. PSNR provides a pixel-level comparison, SSIM evaluates structural similarity, and LPIPS gives a perceptual comparison. Since these metrics are computed from rendered images, they do not directly measure geometric accuracy. They are therefore interpreted as image-based indicators of reconstruction quality.

7.3.3 Computational and practical cost

The computational and practical cost of each strategy is evaluated using 3DGS training time and final model size. Training time describes the computational effort required to optimize the 3DGS model, while model size describes the storage requirement of the final reconstruction.
These metrics do not directly measure reconstruction quality. Instead, they are used to evaluate practical trade-offs. The NBV-guided strategy requires intermediate reconstruction and viewpoint planning steps before the final model is trained. This introduces additional computational cost compared with the predefined strategy. The NBV strategy is therefore only considered beneficial if this added complexity leads to improved reconstruction quality, improved coverage in the rendered views, or a more efficient viewpoint sequence.

7.4 Visual comparison

In addition to the quantitative metrics, the final reconstructions are inspected visually from the same evaluation viewpoints. This comparison is used to identify visible differences that may not be fully captured by PSNR, SSIM, or LPIPS.
The visual inspection focuses on whether the reconstructed object appears complete and stable across viewpoints. Particular attention is given to missing regions, floating artifacts, blurred or unstable surfaces, and thin structures. Since no ground-truth geometry is available, this visual comparison is treated as a qualitative complement to the quantitative image-based metrics rather than as a direct measurement of geometric accuracy.

7.5 Results

The results are obtained using the controlled proxy setup described in Section 7.2. The predefined and NBV strategies are therefore evaluated with the same reconstruction pipeline, the same total number of images, and the same evaluation views.

7.5.1 Image registration and reconstruction information

Table 7.4 summarizes the number of acquired and registered images for each strategy. The table also reports the number of sparse COLMAP points and the number of optimized 3D Gaussians, which are used as practical indicators of reconstruction density.
For the office chair with box, the batch predefined 30,000-iteration reconstruction registered all 36 training images, while the batch NBV reconstruction registered 33 out of 36 images. In the staged reconstructions, the predefined reconstruction registered 35 images and the staged NBV reconstruction registered 34 images. This means that the predefined strategy registered slightly more images in both batched and staged reconstructions.
At the same time, the NBV reconstructions produced substantially denser reconstructions in terms of sparse points and Gaussians. This is especially clear for the

Table 7.4: Image registration and reconstruction density results.

Object	Strategy	Images	Reg.	Gaussians	Sparse points
Office chair with box	Batch Predefined	36	36	50,779	10,312
	Batch NBV	36	33	199,335	19,604
	Stage-wise Predefined	36	35	47,236	9,973
	Stage-wise NBV	36	34	79,196	19,774
Quadrotor	Batch Predefined	36	36	41,592	32,192
	Batch NBV	36	36	48,452	23,690
	Stage-wise Predefined	36	36	25,075	34,531
	Stage-wise NBV	36	36	35,165	22,165

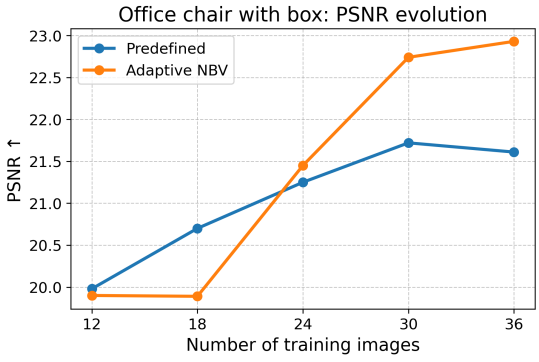
batch reconstruction, where the NBV strategy produced 199,335 Gaussians compared with 50,779 for the predefined strategy. The same trend is seen in the staged comparison, where the staged NBV reconstruction produced 79,196 Gaussians and 19,774 sparse points, compared with 47,236 Gaussians and 9,973 sparse points for the staged predefined reconstruction.
For the quadrotor, all strategies registered all of the images. The predefined strategies registered more sparse points while the NBV strategies produced more Gaussians for the respective batched and staged strategies.
The registration results indicate how much of the acquired data could be used by the reconstruction pipeline. If one strategy has a lower registration count, this must be considered when interpreting the reconstruction-quality results. However, the amount of usable reconstruction information is not determined only by the number of registered images. It also depends on image quality, viewpoint distribution around the object, and overlap between neighbouring views. The density values should therefore be interpreted together with the image-based metrics and visual comparison. A denser reconstruction can indicate improved coverage or more reconstructed detail, but it can also increase model size and does not by itself prove higher reconstruction quality.

7.5.2 Image-based reconstruction quality

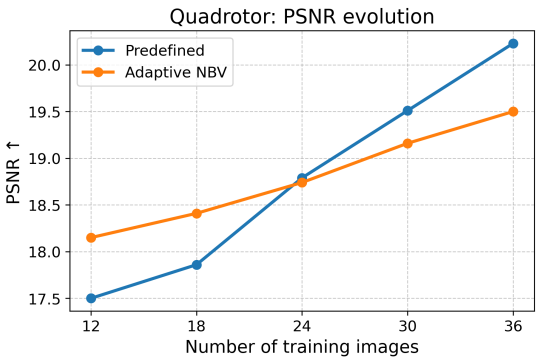
The predefined and NBV strategies are compared at 12, 18, 24, 30, and 36 training images using the same held-out evaluation views, see Table 7.5. The metrics evolution over the stages can be seen in Figures 7.1, 7.2 and 7.3 for the PSNR, SSIM and LPIPS metrics, respectively for both evaluated objects. The result for the batch reconstructions for both evaluated objects can be seen in Table 7.6.
For the office chair with box, the staged results show that the predefined strategy performs slightly better during the earliest stages. At 12 and 18 images, the predefined strategy achieves higher PSNR and SSIM and lower LPIPS. After more NBV-guided views are added, the NBV strategy improves more strongly. At 30 and 36 images, the staged NBV reconstruction achieves higher PSNR and lower LPIPS than the staged predefined reconstruction, while SSIM remains very similar.
The batch comparison shows a similar trade-off for the office chair with box. The NBV reconstruction achieves better PSNR and LPIPS, with a PSNR of 23.89 compared with 22.45 and an LPIPS of 0.191 compared with 0.212. However, the pre-

Table 7.5: Stage-wise image-based reconstruction quality results. Each pair of rows compares the predefined and NBV strategies using the same number of training images. Bold values indicate the better result within each predefined/NBV pair.

Object	Stage-wise Strategy	Images	PSNR ↑	SSIM ↑	LPIPS ↓
Office chair with box	Predefined	12	**19.98**	**0.783**	**0.244**
	NBV	12	19.90	0.781	0.246
	Predefined	18	**20.70**	**0.795**	**0.240**
	NBV	18	19.89	0.775	0.249
	Predefined	24	21.25	**0.810**	**0.229**
	NBV	24	**21.45**	0.800	**0.229**
	Predefined	30	21.72	0.814	0.226
	NBV	30	**22.74**	**0.821**	**0.212**
	Predefined	36	21.61	**0.823**	0.220
	NBV	36	**22.93**	0.822	**0.205**
Quadrotor	Predefined	12	17.50	0.761	0.211
	NBV	12	**18.15**	**0.785**	**0.188**
	Predefined	18	17.86	0.775	0.199
	NBV	18	**18.41**	**0.793**	**0.180**
	Predefined	24	**18.79**	0.792	0.185
	NBV	24	18.74	**0.802**	**0.170**
	Predefined	30	**19.51**	0.808	0.172
	NBV	30	19.16	**0.811**	**0.161**
	Predefined	36	**20.23**	**0.822**	0.159
	NBV	36	19.50	0.817	**0.156**

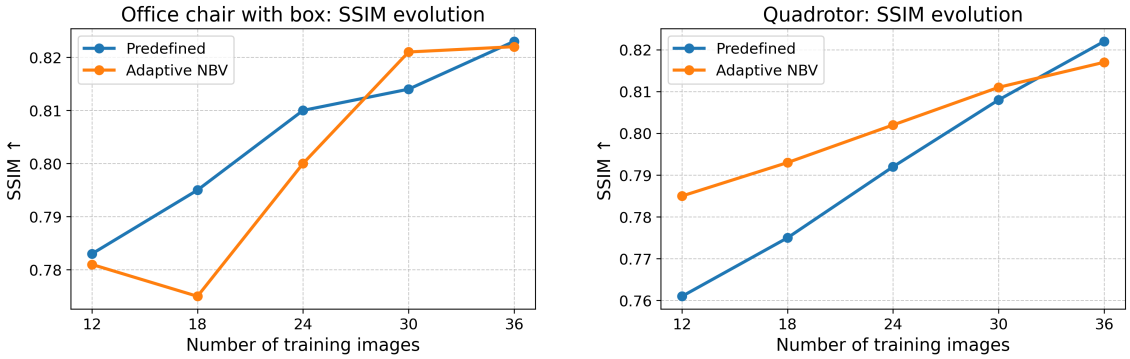


(a) Office chair with box PSNR evolution



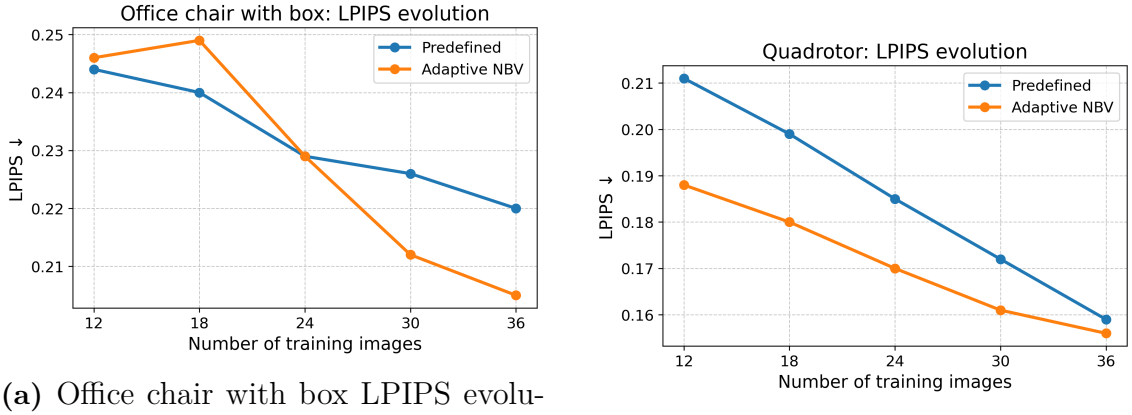
(b) Quadrotor PSNR evolution

Figure 7.1: Evolution of PSNR for the predefined and NBV strategies.



(a) Office chair with box SSIM evolution **(b)** Quadrotor SSIM evolution

Figure 7.2: Evolution of SSIM for the predefined and NBV strategies.



(a) Office chair with box LPIPS evolution

(b) Quadrotor LPIPS evolution

Figure 7.3: Evolution of LPIPS for the predefined and NBV strategies.

Table 7.6: Final image-based reconstruction quality results at 36 training images. Bold values indicate the better result within each predefined/NBV pair.

Object	**Batch Strategy**	**Images**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
Office chair with box	Predefined	36	22.4531	**0.8388**	0.2120
	NBV	36	**23.8903**	0.8133	**0.1910**
Quadrotor	Predefined	36	**21.7588**	**0.8380**	**0.1318**
	NBV	36	21.4060	0.8298	0.1354

defined reconstruction achieves the higher SSIM, with 0.839 compared with 0.813. For the office chair with box, the NBV strategy therefore improves pixel-wise and perceptual image quality, but it does not improve all image-based metrics.
For the quadrotor, the staged results for the NBV strategy does not indicate a clear improvement over the predefined strategy. The NBV strategy starts ahead in all metrics but is overtaken by the predefined strategy in the metric PSNR after 24 images and SSIM after all registered images. The NBV strategy does achieve a lower LPIPS value, but the evolution shows that the predefined strategy decreases at a faster rate. Since both strategies use the same initial 12 images, their Stage 1 metric values should theoretically be identical; the small observed differences are attributed to non determinism in the 3DGS training process. Nevertheless, the predefined strategy improves more rapidly and more consistently across stages.
The batch comparison for the quadrotor shows that the predefined strategy outperforms the NBV strategy in all metrics.
The results should be interpreted together with the visual comparison of rendered images. This is important because image-based metrics may not fully capture visible reconstruction defects such as floaters, holes, blurred regions, or unstable geometry across viewpoints.

7.5.3 Computational and practical cost

Table 7.7 summarizes the practical reconstruction cost for each strategy. The batch reconstructions are trained for 30,000 iterations, while the staged reconstructions represent the cumulative staged process from 30,000 to 70,000 iterations. Training time and model size are recorded to evaluate the practical cost of each approach.

Table 7.7: Computational reconstruction results.

Object	Strategy	Iterations	Time [min]	Model size [MB]
Office chair with box	Batch Predefined	30,000	32.7	12.59
	Batch NBV	30,000	39.3	49.44
	Stage-wise Predefined	70,000	97.3	11.72
	Stage-wise NBV	70,000	102.0	19.64
Quadrotor	Batch Predefined	30,000	16.9	10.32
	Batch NBV	30,000	19.8	12.02
	Stage-wise Predefined	70,000	57.5	6.22
	Stage-wise NBV	70,000	59.7	8.72

For the office chair with box, the batch NBV reconstruction required more training time and produced a substantially larger model than the predefined reconstruction. The batch NBV reconstruction training required 39.3 minutes and resulted in a 49.44 MB model, while the batch predefined reconstruction training required 32.7 minutes and resulted in a 12.59 MB model. The staged NBV reconstruction also required slightly more time and storage than the staged predefined reconstruction, with 102.0 minutes and 19.64 MB compared with 97.3 minutes and 11.72 MB.
This shows that the NBV strategy introduces additional practical cost, for the office chair with box. This added cost is accompanied by improved PSNR and LPIPS, but

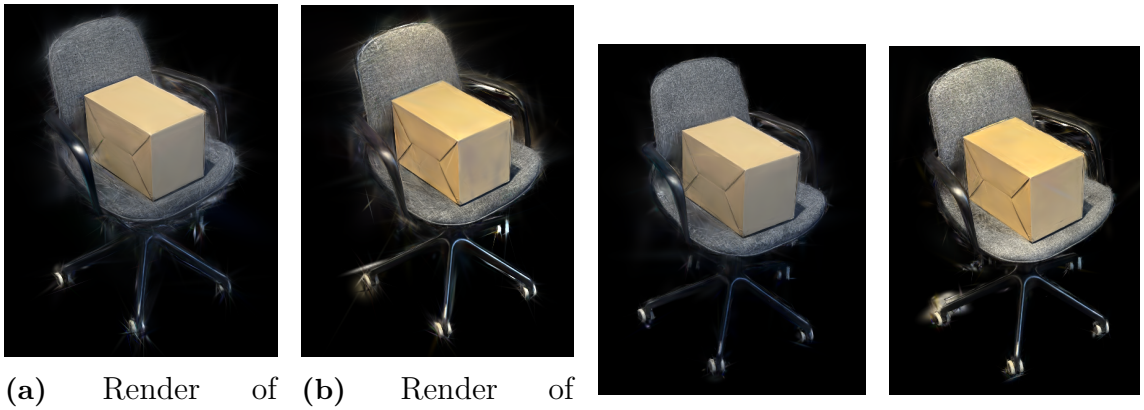
not by improved SSIM. The benefit of the NBV strategy is therefore not a uniform improvement across all metrics, but rather a trade-off between improved perceptual and pixel-wise reconstruction quality, denser reconstruction, and increased computational and storage cost.
For the quadrotor the data shows a similar pattern in terms of time and model size. The pattern is although not as strong as for the chair with box. In this case the additional practical cost, is only accompanied by improved LPIPS in the staged comparison.

7.5.4 Visual comparison

The visual comparison does not provide quantitative measurements, but instead evaluates qualitative differences between the reconstructed 3DGS models. The first section evaluates the office chair with box and the second section evaluates the quadrotor.

7.5.4.1 Renders of Office Chair with Box

Figure 7.4 presents rendered views of the four reconstruction strategies for the office chair with box. All renders are captured from the same viewpoint to ensure a fair comparison between the models. The viewpoint presents the object from a slightly elevated side angle. Figure 7.5 presents the same models from a rear viewpoint.



(a) Render of Stage-wise Predefined model

(b) Render of Stage-wise NBV model

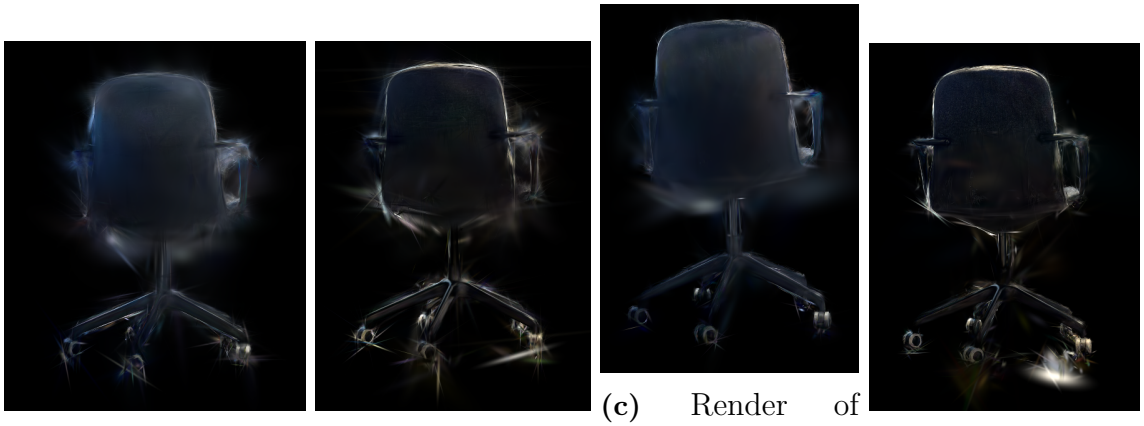
(c) Render of Batch Predefined model

(d) Render of Batch NBV model

Figure 7.4: Rendered images of each 3DGS model from the four different strategies, captured from the same above and side view in each 3DGS model.

The rendered images appear visually similar overall, although several differences can be observed. The staged 3DGS models in Figures 7.4a and 7.4b show more leakage of Gaussians into the background, resulting in lower contrast between foreground and background compared to the batched 3DGS models in Figures 7.4c and 7.4d. This effect is particularly visible along the upper edges of the 3DGS models, where the boundary between foreground and background appears more blurred. The 3DGS reconstruction of the wheels is challenging for most strategies. Among the four models, the batch predefined 3DGS model in Figure 7.4c preserves the wheel structures

most clearly, while the batch NBV 3DGS model in Figure 7.4d shows the largest amount of distortion and Gaussian leakage in these regions. A noticeable difference in brightness and contrast can be observed between the predefined and NBV 3DGS models. The NBV models display stronger contrast on the side surfaces of the box, which may indicate improved surface definition. However, this difference is likely influenced by variations in environmental lighting during data collection rather than solely by the reconstruction strategy itself.
Overall, the batch reconstructions produce visually cleaner results due to their sharper separation between foreground and background. Between the two batch approaches, the NBV reconstruction achieves stronger contrast, while the predefined reconstruction appears more stable and exhibits less Gaussian leakage, particularly around the wheels.



(a) Render of Stage-wise Predefined model

(b) Render of Stage-wise NBV model

(c) Render of Batch Predefined model

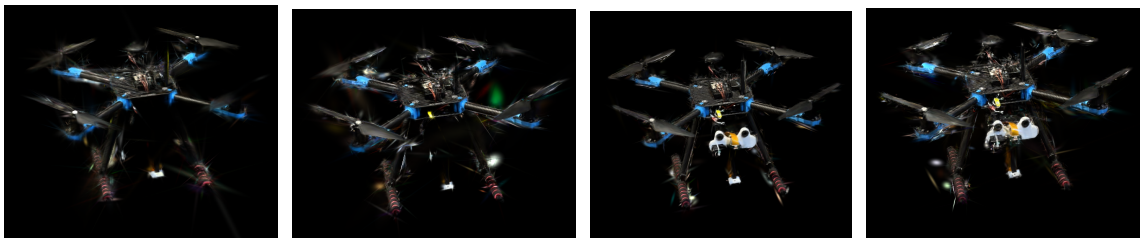
(d) Render of Batch NBV model

Figure 7.5: Rendered images of each 3DGS model from the four different strategies, captured from the same back view in each 3DGS model.

Compared to the previous viewpoint, the rendered images from the rear view exhibit more noticeable visual differences between the reconstruction strategies. From the rear, the NBV-guided 3DGS models show stronger contrast around the object boundaries, particularly in the region beneath the seat. The surface details on the object also appear more textured on the backside of the chair in the NBV-guided 3DGS models in comparison to the predefined 3DGS models which appear more blurred. From the rear it is further clear that the batch predefined model in Figure 7.5c looks the best around the wheels, meanwhile the batch NBV model in Figure 7.5d shows the largest amount of distortion and Gaussian leakage in these regions. Overall the NBV-guided 3DGS models in Figures 7.5b and 7.5d provide the most visually detailed results from the rear viewpoint. Between these two, the batch model exhibits less Gaussian leakage across most regions of the model and maintains sharper boundaries between foreground and background.

7.5.4.2 Quadrotor

It is visually clear that the batch 3DGS models portray the quadrotor better than the staged models. The batched ones has captured more detail and has less leakage of



(a) Render of Predefined & Staged **(b)** Render of NBV & Staged **(c)** Render of Predefined & Batch **(d)** Render of NBV & Batch

Figure 7.6: Rendered images of each 3DGS model from the four different strategies.

Gaussians between foreground and background. It is clear that all models except the batch predefined model in Figure 7.6c has been severely punished by the background masking step. This can be observed in the staged models in Figures 7.6a and 7.6b where the white frame in the front holding the servo and capture camera is gone. The front right propeller is vague in the staged NBV model in Figure 7.6b and the back left propeller is half missing in the batch NBV model in Figure 7.6d.
Overall the batch predefined model in Figure 7.6c provides the most visually clean result. It has the lowest amount of distortion and Gaussian leakage, and also contains most details about the quadrotor. Although the model in Figure 7.6c contains the most detail, it has to do with the result of the masking of the images and less to do with the 3DGS training. However, all images are masked using the same model and therefore treated fairly. The masking of the images is affected by the viewpoint of the quadrotor, and therefore directly influenced by the NBV proposals.

7.6 Quadrotor PoC: practical limitations

Several practical limitations were identified during implementation and testing of the quadrotor system. These limitations affected the reliability of the intended autonomous workflow and motivated the use of a controlled proxy setup for the final strategy comparison. The main limitations are summarized in Table 7.8.
These limitations do not invalidate the quadrotor implementation, but they define the maturity level of the system. The PoC demonstrates that the main components can be connected, while also showing that additional work is required before the complete workflow can be executed autonomously and repeatably indoors.

7.7 Summary of evaluation

The evaluation compares a predefined image acquisition strategy with an NBV image acquisition strategy. The predefined strategy acquires 36 images from a fixed circular viewpoint pattern over three height levels. The NBV strategy starts from 12 initial images and adds 24 images through repeated NBV planning, resulting in the same final number of images.
The comparison is performed using the controlled proxy setup rather than the

Table 7.8: Practical limitations of the quadrotor PoC system.

Issue	**Observed effect and consequence**
Limited indoor GPS reliability	GPS was not reliable as a primary indoor position source, which made repeatable autonomous indoor flight difficult.
ArUco visibility limitations	Markers were not always visible due to viewpoint, distance, or FOV limitations, causing intermittent pose updates.
Vibration and motion blur	Propeller and frame vibrations reduced image sharpness, making captured images less reliable for 3DGS training.
Lighting sensitivity	Glare and exposure variations affected marker detection and image quality, reducing robustness across test conditions.
Off-board 3DGS computation	3DGS training required external GPU resources, so the full acquisition—reconstruction—planning loop could not run onboard in real time.
Closed-loop robustness	The full sequence from flight to NBV-guided reacquisition was not repeatable enough for final quantitative evaluation.

quadrotor PoC system. This allows the acquisition strategies to be evaluated under controlled and repeatable conditions. The practical limitations of the quadrotor system are presented in Section 7.6.
The evaluation uses image registration, PSNR, SSIM, LPIPS, sparse COLMAP point count, number of Gaussians, training time, model size, and visual inspection. Metrics that require ground-truth geometry are excluded because no exact reference geometry is available. The results are therefore interpreted as a comparison of image-based reconstruction quality, reconstruction density, and practical cost rather than as an absolute measurement of geometric accuracy.
For the office chair with box, the NBV strategy improves on PSNR and LPIPS, while the predefined strategy achieves slightly higher SSIM, in both the batch and stage-wise evaluations. The NBV reconstructions produce denser models, but with higher computational and storage cost. The final conclusion for this object therefore depends on whether perceptual quality and reconstruction density are considered more important than compactness, runtime, and structural-similarity score.
For the quadrotor, the NBV strategy creates denser models and has higher computational and storage cost. This is only accompanied by a better LPIPS value for the staged strategy but not the batched strategy, in comparison to the respective predefined models. The metrics PSNR and SSIM are higher in both the staged and the batch results for the predefined models.

8

Discussion

This chapter discusses the results and limitations of the thesis. The discussion is divided into four main parts. First, the change from the intended quadrotor-based image acquisition system to the controlled proxy setup is discussed. Second, research questions are addressed based on the evaluation results presented in Chapter 7. Third, directions for future work are outlined. Finally, the overall conclusions of the thesis are presented.

8.1 Changing the image capturing system

The thesis was originally intended to use the quadrotor-based PoC system for the final image acquisition experiments. Although the quadrotor was not expected to be a fully autonomous product-level system, it needed to be sufficiently reliable for repeatable indoor acquisition. During implementation and testing, this was not achieved. The main limiting factor was the reliability of the pose estimation and position control.

The intended approach was to use ArUco-based pose estimation together with the Pixhawk's onboard sensor estimates to control the quadrotor indoors. However, without GPS the quadrotor relied mainly on barometric measurements, IMU data, and intermittent corrections from ArUco marker detections. This resulted in unstable vertical and horizontal behaviour. The take-off altitude varied between trials, and the quadrotor often showed a gradual altitude drift after take-off. Horizontal drift was also observed, which caused the tracking camera to lose sight of the ArUco markers. Once the markers were no longer visible, the system could not provide reliable pose corrections.

A denser ArUco marker layout could potentially reduce this problem, but it would not be a practical solution. A large number of markers would have to be placed accurately before each experiment, which would make the setup time-consuming and reduce the benefit of autonomous acquisition. Outdoor operation with GPS was also tested briefly, even though it was outside the main scope of the thesis. These tests showed that wind, vibration, GPS uncertainty, and changing lighting conditions introduced additional problems. In particular, vibration and sunlight reduced image quality, while sunlight and glare made ArUco marker detection less reliable.

For these reasons, the final comparison between the predefined and NBV acquisition strategies was performed using the controlled proxy setup instead of the quadrotor. The proxy setup still preserved the central parts of the intended workflow: a tracking

camera was used for pose estimation, a separate capture camera was used for image acquisition, and the acquired images were processed through the same reconstruction and NBV planning pipeline.
This change means that the final evaluation should be interpreted as a controlled proxy evaluation of the acquisition strategies rather than as a full demonstration of autonomous quadrotor acquisition. The results still address whether NBV-guided viewpoint selection can improve 3DGS reconstruction, but it does not prove that the implemented quadrotor can execute the complete workflow autonomously.

8.2 Discussion of Research Questions

The thesis was guided by three research questions. The first two concern the comparison between predefined and NBV-guided image acquisition. The third concerns the practical limitations of the implemented quadrotor PoC system.

- **RQ1**: How does image-based reconstruction quality and visual completeness compare between an adaptive NBV-based and a predefined image acquisition strategy when using the same number of images?
- **RQ2**: How does the number of images required to reach comparable image-based reconstruction quality and visual completeness differ between an adaptive NBV-based and a predefined image acquisition strategy in a controlled proxy setup?
- **RQ3**: What practical limitations prevent the implemented quadrotor PoC system from completing the same acquisition and reconstruction workflow autonomously in an indoor environment?

The following sections discuss each research question based on the implementation and evaluation results.

RQ1: Reconstruction quality and completeness using the same number of images

For the office chair with a box, the staged results show that the NBV strategy does not consistently outperform the predefined strategy in the early acquisition stages. At 12 and 18 images, the predefined strategy achieves slightly higher PSNR and SSIM and lower LPIPS. The NBV strategy begins to outperform the predefined strategy from 30 images onward, achieving higher PSNR and lower LPIPS, while SSIM remains comparable. The batch comparison confirms this pattern: the NBV strategy achieves higher PSNR and lower LPIPS, while the predefined strategy achieves higher SSIM. The visual comparison supports a mixed interpretation: the NBV reconstructions show subtly stronger contrast and surface detail on both the rear of the chair and the front surfaces of the box, while the predefined reconstructions are slightly more stable around the wheels.
For the quadrotor, the staged results show that the NBV strategy leads on all three metrics at 12 and 18 images, but the predefined strategy overtakes on PSNR after 24 images and on SSIM across most later stages. The NBV strategy retains a lower LPIPS value throughout the staged comparison. The batch comparison shows that

the predefined strategy outperforms the NBV strategy on all three metrics. The visual comparison supports this: the predefined batch reconstruction produces the cleanest overall result, with less Gaussian leakage and the most complete rendering of structural details.
Taken together, the results show that image-based reconstruction quality and visual completeness differ inconsistently between the two strategies when using the same number of images. The NBV strategy does not uniformly improve image-based reconstruction quality over the predefined strategy for the same number of images. The improvement is most consistent for PSNR and LPIPS on the chair, while SSIM favours the predefined strategy in the final comparison for both objects. This suggests that the benefit of NBV-guided image acquisition is not guaranteed and may depend on object geometry and how well the initial reconstruction captures the object before NBV selection begins.

RQ2: Reconstruction quality relative to acquisition effort

For the office chair with a box, the staged comparison provides some evidence of an efficiency gain consistent with RQ2. The NBV strategy at 30 images achieves a higher PSNR and lower LPIPS than the predefined strategy at 36 images, while SSIM is nearly equal. This indicates that the NBV strategy can reach comparable or better reconstruction quality on these metrics with fewer acquired images, at least for this object. However, this efficiency comes at a higher computational and storage cost, as shown in Table 7.7, and the improvement does not extend consistently to all metrics.
For the quadrotor, the staged data does not support a comparable efficiency gain. The NBV strategy leads on all three metrics at 12 and 18 images and retains a lower LPIPS value throughout, but the predefined strategy achieves higher PSNR and SSIM from 30 images onward. The NBV strategy therefore does not reach comparable overall quality with fewer acquired images for this object, and it is not possible to confirm an efficiency gain based on the available staged results.
Overall, the results show that the number of images required to reach comparable quality differs inconsistently between the two strategies. For the office chair with a box, the NBV strategy reaches comparable or better quality on PSNR and LPIPS at 30 images compared with the predefined strategy at 36, suggesting a modest efficiency gain. This benefit is not confirmed for the quadrotor. The acquisition efficiency of the NBV strategy must also be weighed against the additional computational cost introduced by the intermediate reconstruction and NBV planning steps, which are required before each new set of images are acquired.

Interpretation of the NBV results

The results from RQ1 and RQ2 show that the NBV strategy did not consistently outperform the predefined strategy across both objects, metrics, and acquisition stages. Several factors may have contributed to this behaviour, although their individual effects cannot be determined from the current evaluation.
The results were therefore less uniformly positive for NBV than expected, since

the NBV strategy was expected to provide clearer benefits by targeting under-represented regions of the reconstruction. However, the mixed results also indicate that identifying informative viewpoints is not sufficient on its own; the selected viewpoints must also be suitable for the complete reconstruction pipeline.

This section provides possible interpretations of the observed trends, while noting that these explanations cannot be confirmed without additional experiments.

One possible explanation is that the outcome of the NBV strategy depends strongly on the intermediate reconstruction used for planning. The purpose of NBV is to identify under-represented or uncertain regions in the current model and select viewpoints that can improve those regions. However, the reconstruction and training pipeline contains non-deterministic components, which means that the exact intermediate reconstruction may vary between runs even when the same images are used. As a result, the regions considered under-represented by the NBV planner may also vary. A different iteration of the same experiment could therefore have produced slightly different NBV selections and possibly different reconstruction results. Since the evaluation was limited to one run per object, it is not possible to determine how repeatable the observed differences are.

Another possible explanation concerns the formulation of the NBV evaluation score. The implemented score is based on a pixel-wise sum of the rendered information gain. As a result, viewpoints where the object occupies more image pixels may receive a higher score than viewpoints where the object appears smaller. In the experiments, candidate viewpoints were generated on a cylindrical surface around the object, which helps make candidates at similar heights more comparable. However, for the quadrotor object, which has a relatively low height compared with its horizontal extent, candidate viewpoints from higher elevations may have been farther from parts of the object than side views. These views could potentially have provided additional information from above, but may have received a lower score because the object occupied fewer pixels in the rendered view. This cannot be confirmed from the current evaluation, but it is a possible explanation for why the NBV strategy may not have selected some views that would have improved coverage of the quadrotor geometry.

The preprocessing step may also have influenced the comparison. In particular, the background masking could have affected the two strategies and objects differently. The chair with box consists of larger and more continuous surfaces, while the quadrotor has thinner structures, open regions, and more complex geometry. If the mask removed parts of the quadrotor structure, introduced noisy object boundaries, or produced inconsistent silhouettes between views, this could have affected both COLMAP registration and the 3DGS reconstructions. This may be one reason why the NBV strategy performed less consistently for the quadrotor. However, since the masking quality was not evaluated separately, its effect cannot be isolated from the other parts of the pipeline.

The results also suggest that the NBV planner could benefit from considering the requirements of image registration more directly. Images that are not registered by COLMAP are not used during 3DGS training. In many cases, this works as a useful protection mechanism, since images that cannot be matched may have insufficient overlap, poor quality, or too few reliable features. However, this can also be a

limitation for NBV-based acquisition. A viewpoint may be valuable according to the NBV score because it observes an under-represented part of the object, but if the image does not match sufficiently with the existing image set, the additional information is lost because the image is excluded from training. In the current evaluation, all quadrotor images were registered, while this was not the case for the chair with box. This means that registration failure does not explain the weaker final NBV result for the quadrotor, but it remains an important consideration for the general robustness of the NBV pipeline.
Overall, the inconsistent results are likely not caused by a single factor. Instead, they may result from the interaction between object geometry, masking quality, intermediate reconstruction variability, NBV score formulation, and COLMAP image matching. The evaluation therefore suggests that NBV-guided acquisition can be beneficial, but that selecting informative viewpoints is not enough on its own. The selected images must also be suitable for masking, COLMAP registration, and 3DGS training in order to consistently improve the final reconstruction.

RQ3: Practical limitations of the quadrotor PoC system

The quadrotor PoC system was not able to complete the image acquisition and reconstruction workflow autonomously indoors. The limitations are documented in detail in Section 7.6; the discussion here focuses on what they reveal collectively.
The most significant finding is that the limitations are not independent: they form a coupled chain. Indoor operation without GPS meant that the system depended on barometric, IMU, and ArUco-based corrections to maintain position. When the quadrotor drifted the tracking camera lost sight of the ArUco markers, removing the only external pose correction available. Without a reliable pose estimate, corrected waypoints could not be computed, horizontal and vertical drift continued uncorrected, and the system could not recover to a position suitable for repeatable image acquisition. Vibration further degraded both image quality and marker detection, compounding the same cascade. The off-board 3DGS computation requirement added a further practical barrier to closing the full acquisition—reconstruction—planning loop.
The result is that the PoC demonstrated successful integration of the individual components — communication, marker detection, image capture, off-board reconstruction, and NBV planning all functioned in isolation or under controlled conditions — but the complete autonomous workflow was not reliable enough to serve as the basis for the final strategy comparison. This answers RQ3: the primary practical limitation is the lack of a robust indoor localization source. Without it, pose estimation becomes intermittent, flight behaviour becomes unrepeatable, and the downstream acquisition quality is insufficient for a controlled evaluation of image acquisition strategies.

8.3 Limitations of the evaluation

The final evaluation provides a controlled comparison between predefined and NBV image acquisition strategies, but several limitations should be considered when in-

terpreting the results.

First, the evaluation does not demonstrate the complete workflow on the quadrotor platform. The controlled proxy setup was used to obtain repeatable image acquisition, which means that the final results do not include the full effects of quadrotor flight, such as vibration, position-control errors, motion blur, and flight-time constraints. The results therefore evaluate the acquisition strategy and reconstruction pipeline more directly than the complete autonomous quadrotor system.

Second, the evaluation is limited by the controlled proxy setup, which was not a fully optimized studio environment. The image acquisition was performed indoors in an office environment, where lighting conditions, background clutter, reflections, shadows, and available space were not ideal. These factors may have affected image quality, background masking, feature detection, COLMAP registration, and the final 3DGS reconstruction. Since both the predefined and NBV strategies were evaluated under the same general conditions, this does not necessarily explain the differences between the strategies by itself. However, it may have influenced the overall reconstruction quality and may have affected objects differently depending on their geometry, texture, and visibility.

Third, the evaluation does not measure absolute geometric reconstruction accuracy. No ground-truth geometry was available for the evaluated objects, so the thesis does not report absolute surface error or exact surface completeness. The evaluation instead relies on image-based rendering metrics, image registration results, theoretical acquisition effort, and visual inspection. These metrics are useful for comparing the two strategies, but they do not fully describe geometric accuracy.

Fourth, the evaluation does not isolate the contribution of each individual component in the NBV planner. The predefined and NBV strategies are compared as complete acquisition strategies, where viewpoint order and viewpoint distribution differ by design. This is appropriate for evaluating whether NBV-based acquisition improves reconstruction quality under the tested conditions, but it does not show how much each part of the NBV scoring function contributes to the final result.

Fifth, the evaluation is limited to a small number of objects. The results therefore show how the system performed under the tested conditions, but they do not prove general performance across all object types. Objects with different texture, reflectance, shape complexity, or occlusion patterns may affect both SfM registration and 3DGS reconstruction quality.

Sixth, the evaluation is limited to a single test for each object. Since parts of the framework are non-deterministic, multiple iterations should be performed to find out the level of uncertainty in the framework result. By executing more iterations using the same images the results would be more robust and trustworthy. At the moment, the repeatability of the results cannot be confirmed.

Seventh, the evaluation does not isolate possible biases in the NBV scoring and preprocessing steps. The NBV score is not explicitly distance-normalised, and the background masking quality was not evaluated separately. These factors may have influenced the selected viewpoints and reconstruction quality, especially for the quadrotor object with its thinner and more complex geometry. In addition, the NBV planner does not explicitly consider how well a selected viewpoint is expected to match with previous images in COLMAP. These factors may have contributed

to the observed differences between the predefined and NBV strategies, but their individual effects cannot be determined from the current evaluation.

8.4 Future Work

Several parts of the system should be improved before the proposed workflow can become a fully autonomous quadrotor-based acquisition system.
One area for future work is the pose estimation and localization system. The quadrotor-based PoC showed that reliable positioning is needed for repeatable autonomous acquisition. A future system could use Ultra-Wideband localization or a similar beacon-based positioning system to provide a more stable external position estimate. This would reduce the dependence on GPS and make it easier to keep the quadrotor within the intended region.
The image acquisition hardware could also be improved. The quadrotor tests showed that vibration, motion blur, and lighting variations can affect both image quality and ArUco marker detection. Future work could therefore investigate improved camera mounting, vibration damping, shorter exposure times and the use of a global-shutter camera. These improvements would support both pose estimation and the quality of the images used for SfM and 3DGS reconstruction.
Another area for future work is to make the 3DGS reconstruction pipeline more efficient. In this thesis, reconstruction was performed off-board and required a significant amount of computation. A faster reconstruction process would make the adaptive loop more practical, especially if 3DGS models need to be updated repeatedly between acquisition steps.
The NBV planner could also be developed further. Future work could investigate how viewpoint selection can be made faster and more robust, and how NBV poses can be generated in closer connection with the current state of the 3DGS reconstruction. In the long term, real-time or near real-time 3DGS reconstruction combined with NBV planning would make it possible to update the reconstruction and select new viewpoints during the acquisition process.
The closed-loop workflow should also be automated further. In the current implementation, several steps still require manual handling, such as transferring data, running reconstruction, planning new viewpoints, and executing the next acquisition step. A future system could automate the full loop from image capture to reconstruction, NBV planning, and execution of the next selected viewpoint.
Future evaluations should include ground truth models. This would make it possible to evaluate geometric reconstruction error and surface completeness directly, instead of relying mainly on image-based rendering metrics and qualitative inspection.
Future work could also investigate NBV scoring functions that are distance-normalised, more robust to segmentation quality, and constrained by expected image overlap or feature matching in COLMAP.
Finally, the method should be evaluated on a larger and more diverse set of objects. Objects with different shapes, textures, reflective properties, and occlusions would make it possible to better understand when NBV-based acquisition is beneficial and when a predefined acquisition strategy is sufficient. The tests should also be iterated to determine the level of uncertainty between runs, since some parts of the

framework are non-deterministic. With a calculated variance of the iterated results, conclusions about the framework would be more robust.

8.5 Conclusion

This thesis investigated adaptive image acquisition for object-centred 3DGS reconstruction using NBV planning. Two connected aims were pursued: implementing and testing a quadrotor-based PoC system for autonomous image acquisition, and evaluating whether an NBV-guided image acquisition strategy can improve reconstruction quality or acquisition efficiency compared with a predefined strategy in a controlled setup.
The quadrotor-based PoC demonstrates that the main system components — RPi and Pixhawk communication, ArUco-based pose estimation, image capture, off-board 3DGS reconstruction, and NBV planning — can be integrated into a common workflow. However, the system was not sufficiently reliable for repeatable autonomous indoor acquisition. The primary limitation was the absence of a robust indoor localization source, which made the pose estimation intermittent and flight behaviour unpredictable. This caused downstream effects on image quality and prevented the full acquisition—reconstruction—planning loop from being completed reliably. The quadrotor therefore remains a PoC whose individual components were verified, but whose complete workflow requires further development before it can support a fair quantitative evaluation.
The controlled proxy evaluation showed that the NBV strategy does not uniformly outperform the predefined strategy. For the office chair with box, the adaptive strategy achieved higher PSNR and lower LPIPS in the batch comparison, and showed subtly stronger surface detail on the rear of the chair and the front of the box, while the predefined strategy achieved higher SSIM. The staged comparison also suggests that the NBV strategy can reach comparable or better PSNR and LPIPS with fewer acquired images for this object, though this comes at a higher computational and storage cost. For the quadrotor object, the NBV strategy led on most metrics at lower image counts in the staged comparison, but the predefined strategy achieved higher PSNR and SSIM from 30 images onward, and outperformed the NBV strategy on all three metrics in the batch comparison. The results therefore show that image-based reconstruction quality and acquisition efficiency differ inconsistently between the two strategies across the tested objects and metrics: the NBV strategy does not uniformly outperform the predefined strategy.
Overall, the thesis shows that NBV-guided image acquisition is a promising direction for object-centred 3DGS reconstruction, but that its benefit is not guaranteed and depends on the object geometry, the acquisition stage, and the metric considered. Although the results do not show a consistent advantage for NBV strategy, they provide useful insight into the practical conditions under which NBV-based acquisition may or may not improve object-centred 3DGS reconstruction. The results also indicate that NBV viewpoint selection should be evaluated together with the full reconstruction pipeline, since masking quality, image registration, score formulation, and 3DGS training can all affect whether selected viewpoints improve the final reconstruction. A fully autonomous quadrotor-based implementation of this

workflow requires more robust indoor localization, improved image quality under flight conditions, and a more efficient reconstruction pipeline before it can be used reliably.

Bibliography

[1] A. Nazim, M. Kondrát, K. Zidek, and J. Piteľ, "Methodology of object reconstruction by photogrammetry and structured-light scanning for industrial 3d visualisation," *Sensors*, vol. 25, p. 7177, 11 2025.

[2] S. Hatami Gazani, M. Tucsok, I. Mantegh, and H. Najjaran, "Bag of views: An appearance-based approach to next-best-view planning for 3D reconstruction," *arXiv preprint arXiv:2307.05832*, 2023.

[3] Y. Xie, Y. Cai, Y. Zhang, L. Yang, and J. Pan, "Gauss-mi: Gaussian splatting shannon mutual information for active 3d reconstruction," 2025.

[4] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[5] M. Maboudi, M. Homaei, S. Song, S. Malihi, M. Saadatseresht, and M. Gerke, "A review on viewpoints and path planning for uav-based 3-d reconstruction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, p. 5026–5048, 2023.

[6] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, July 2023.

[8] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," 2020.

[11] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe, "Bilateral reference for high-resolution dichotomous image segmentation," *CAAI Artificial Intelligence Research*, p. 9150038, Dec. 2024.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, May 2022.

[14] L. Meier, D. Honegger, and M. Pollefeys, "Px4: A node-based multithreaded open source robotics framework for deeply embedded platforms," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6235–6240, 2015.

[15] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A

Appendix

Source Code Repositories

The implementation code for this thesis is available in two public GitHub repositories:

- **Reconstruction pipeline:** `https://github.com/GustavJanson4/enocson.git`
- **Drone PoC:** `https://github.com/GustavJanson4/WirevisionGG.git`

System Requirements and Environment

Off-board workstation (reconstruction pipeline)

The reconstruction pipeline requires the following environment:

- Operating System: Ubuntu 24.04.4 LTS
- Python: 3.12

Onboard system (drone PoC)

The onboard RPi software requires the following environment:

- Host OS: Raspberry Pi OS (with Ubuntu containerized via Docker)
- ROS 2: Jazzy Jalisco
- C++: C++17

Hardware Specifications

All 3D reconstructions in this thesis were conducted on a workstation with the following specifications:

- CPU: AMD Ryzen 7 5700X (8-core, 4.67 GHz boost)
- GPU: NVIDIA GeForce RTX 2080 Ti (11 GiB VRAM, CUDA 12.1)
- Memory: 31 GiB RAM
- Operating System: Ubuntu 24.04.4 LTS

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se



CHALMERS
UNIVERSITY OF TECHNOLOGY