

Robotic Burst Imaging for Light-Constrained 3D Reconstruction

Ahalya Ravendran

A thesis submitted in fulfillment
of the requirements of the degree of
Doctor of Philosophy



Australian Centre for Field Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney

Submitted February 2023; revised August 2023

Statement of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Ahalya Ravendran

23 August 2023

Authorship Attribution Statement

Parts of this thesis are based on the following:

- [Chapter 3](#) of this thesis is published as [\[1\]](#),
A. Ravendran, M. Bryson, and D. G. Dansereau, “Burst imaging for light-constrained structure-from-motion,” *IEEE Robotics and Automation Letters (RA-L, ICRA)*, vol. 7, no. 2, pp. 1040–1047, Apr. 2022.
I designed the study, analyzed the data and wrote the drafts of the manuscript.
- [Chapter 4](#) of this thesis is presently under submission and is available as a preprint [\[2\]](#).
I designed the study, analyzed the data and wrote the drafts of the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Ahalya Ravendran

23 August 2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Dr. Donald Dansereau

23 August 2023

Abstract

Robotic Burst Imaging for Light-Constrained 3D Reconstruction

This thesis proposes a novel input scheme, robotic burst, to improve vision-based 3D reconstruction for robots operating in low-light conditions, where existing state-of-the-art robotic vision algorithms struggle due to low signal-to-noise ratio in low-light images. We aim to improve the correspondence search stage of feature-based reconstruction using robotic burst imaging, including burst-merged images, a burst feature finder, and an end-to-end learning-based feature extractor. Firstly, we establish the use of robotic burst imaging to compute burst-merged images for feature-based reconstruction. We then develop a burst feature finder that locates features with well-defined scale and apparent motion on a burst to deal with limitations of burst-merged images such as misalignment at strong noise. To improve feature matches in burst-based reconstruction, we also present an end-to-end learning-based feature extractor that finds well-defined scale features directly on light-constrained bursts.

We evaluate our methods against state-of-the-art reconstruction methods for conventional imaging that uses both classical and learning-based feature extractors. We validate our novel input scheme using burst imagery captured on a robotic arm and drones. We demonstrate progressive improvements in low-light reconstruction using our burst-based methods against conventional approaches and overall, converging 90% of all scenes captured in millilux conditions that otherwise converge with 10% success rate using conventional methods. This work opens up new avenues for applications, including autonomous driving and drone delivery at night, mining, and behavioral studies on nocturnal animals.

Acknowledgements

I would like to take this opportunity to express my heartfelt gratitude to all the individuals who have inspired and supported me throughout my doctoral research.

I am deeply grateful to my supervisor, Dr. Donald Dansereau, for his unwavering support, encouragement, and exceptional mentorship throughout my research journey. His intuitive thinking, profound knowledge, and insightful feedback have helped me develop both as a researcher and an individual, and I will always cherish the memories of working under his supervision.

I would like to extend my gratitude to my co-supervisor, Dr. Mitch Bryson. His valuable feedback and constant motivation have been instrumental in shaping my research. I would also like to express my appreciation to my thesis committee members, Dr. Virole Ila and Dr. Oscar Pizzaro, for their supportive recommendations and constructive criticism, which played a great role in enhancing the quality of work.

I am privileged to have been a part of the Robotic Imaging group. Heartfelt thanks to Ryan Griffiths, Jack Naylor, Bhargava Gowda, Chengyang Yan, Nikolai Goncharov, and Denash Khan for all the engaging questions over the years, brainstorming sessions, and most of all, for the laughter we shared together. Special thanks to Dr. Dorian Tsai, Dr. Stephany Berrio Perez, and Dr. Tejaswi Digumarti for sharing their academic experience and valuable time in providing beneficial suggestions.

I would like to express my sincere gratitude to all the researchers at the Australian Centre for Field Robotics for creating a supportive and intellectually stimulating environment. I am grateful for all the valuable discussions and insights, which have broadened my knowledge and helped me gain a deeper understanding. With great appreciation, I also express my thanks to the administrative staff for looking out for me, especially during the pandemic, in ensuring my well-being.

I would like to acknowledge the Engineering and Information Technologies Research Scholarship from the University of Sydney for providing me with financial support throughout my research. Their generosity has enabled me to pursue my research goals and contribute to the scientific community.

Lastly, I would like to express my gratitude to my family for giving me a chance at life and for being an unwavering source of comfort throughout my journey.

*Robots, in sufficient quantity and given enough time,
will end up writing, reviewing and hopefully citing this thesis.*

Contents

Statement of Originality	i
Authorship Attribution Statement	ii
Abstract	iii
Acknowledgements	iv
Contents	vi
List of Figures	x
List of Tables	xii
Nomenclature	xiii
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	4
1.3 Contributions	5
1.4 Significance	6
1.5 Outline	6

2	Background	9
2.1	Overview	9
2.2	3D Reconstruction	10
2.2.1	SfM in Detail	13
2.3	Imaging Trade-offs	16
2.4	Burst Imaging	19
2.5	Noise Modeling	21
3	Light-Constrained Burst SfM	24
3.1	Overview	24
3.2	Related Work	26
3.3	Robotic Burst-Based Structure-from-Motion	29
3.3.1	Image Acquisition for Reconstruction	29
3.3.2	Robotic Burst Alignment	30
3.3.3	Robust Temporal Merge	32
3.4	Results	34
3.4.1	Feature Performance in Noise	34
3.4.2	Reconstruction Performance	37
3.4.3	Camera Trajectory Accuracy	39
3.4.4	Speed	41
3.5	Discussion and Future Directions	43
4	Burst Feature Finder	46
4.1	Overview	46
4.2	Related Work	48
4.3	Robotic Burst Feature Finder	49
4.3.1	Apparent Motion within a Robotic Burst	50
4.3.2	Motion Stack	51
4.3.3	Scale Slope Search Space	53

4.3.4	Descriptor	53
4.3.5	Computational Complexity	54
4.4	Results	54
4.4.1	Feature Performance in Noise	55
4.4.2	Reconstruction Performance	58
4.4.3	Camera Trajectory Accuracy	62
4.4.4	Speed	63
4.4.5	Failure Cases	65
4.5	Discussion and Future Directions	66
5	Learned Burst Feature Finder	67
5.1	Overview	67
5.2	Related Work	70
5.3	Learned Robotic Burst Feature Finder	72
5.3.1	Joint Learning of Detection and Description	73
5.3.2	Robotic Burst Generation for Training	75
5.3.3	Robotic Burst Generation for Inference	77
5.4	Results	77
5.4.1	Implementation Details	77
5.4.2	Feature Performance in Noise	78
5.4.3	Reconstruction Performance	82
5.4.4	Drone-based Validation	84
5.4.5	Speed	90
5.4.6	Ablation Study	92
5.4.7	Failure Cases	93
5.5	Discussion and Future Directions	95

6	Conclusions and Future Directions	96
6.1	Conclusions	96
6.2	Future Directions	98
6.2.1	Adaptive Robotic Burst Imaging	98
6.2.2	Raw Robotic Burst Imaging	99
6.2.3	Sensor Fusion	100
6.2.4	Broader Impact	101
	Bibliography	102

List of Figures

1.1	An example light-constrained scene	2
2.1	Conceptual visualisation of feature-based SfM	12
2.2	Conceptual visualisation of an example SfM pipeline	14
2.3	A gold standard image in a well-lit condition	17
2.4	A naïve image captured in low light	17
2.5	An image captured in low light with wide aperture	18
2.6	An image captured in low light with long exposure	18
2.7	An image captured in low light with high gain	19
2.8	An example burst photography capture	19
2.9	An example burst captured in low light	20
2.10	Noise model of captured imagery	21
3.1	Summarised visualisation of robotic burst-based SfM	26
3.2	Alternative exposure scheme for robotic burst-based SfM	27
3.3	Robotic burst-based SfM architecture	30
3.4	Pairwise hierarchical tile alignment	31
3.5	Direct method within a robotic burst	33
3.6	Alternative exposure schemes of captured imagery	33
3.7	Feature performance of burst-based SfM	35
3.8	Noise performance of robotic burst-based SfM	36
3.9	Captured dataset for reconstruction evaluation	37

3.10	Camera pose estimation for robotic burst-based SfM	41
4.1	Summarised visualisation of robotic burst feature finder	47
4.2	Robotic burst feature finder for reconstruction	50
4.3	Local motion in a captured robotic burst	51
4.4	Robotic burst feature finder architecture	52
4.5	Robotic burst feature finder performance in noise	55
4.6	ROC curve for robotic burst feature finder	56
4.7	Impact of images in robotic burst feature performance	57
4.8	Captured motion-driven dataset for reconstruction evaluation	59
4.9	Camera pose estimation for robotic burst feature finder	64
4.10	Failure case examples for robotic burst feature finder	65
5.1	Summarised visualisation of learned robotic burst feature finder	69
5.2	State-of-the-art feature detection on a light-constrained image	72
5.3	Learned robotic burst features for reconstruction	73
5.4	Learned robotic burst feature finder architecture	74
5.5	Learned robotic burst generation pipeline	76
5.6	Learned robotic burst feature matching performance in noise	79
5.7	Learned robotic burst feature performance in noise	80
5.8	Camera pose estimation for learned robotic burst feature finder	81
5.9	Captured drone imagery dataset for reconstruction evaluation	85
5.10	Learned robotic burst feature performance on drone imagery	86
5.11	Learned robotic burst feature finder performance	86
5.12	Overall feature performance for light-constrained reconstruction	87
5.13	Overall reconstruction performance for a light-constrained scene	91
5.14	Ablation study on keypoint variation	92
5.15	Ablation study on patch size variation	93
5.16	Noise limitation of learned robotic burst feature finder	94
5.17	Failure case examples for learned robotic burst feature finder	94

List of Tables

3.1	Reconstruction performance for robotic burst-based SfM	40
3.2	Camera pose estimation for robotic burst-based SfM	42
4.1	Reconstruction performance for robotic burst feature finder	61
4.2	Camera pose estimation for robotic burst feature finder	62
4.3	Speed performance for robotic burst feature finder	63
5.1	Feature matching performance for learned robotic burst feature finder	81
5.2	Camera pose estimation for learned robotic burst feature finder . . .	82
5.3	Reconstruction performance for learned robotic burst feature finder .	83
5.4	Camera pose estimation for learned robotic burst feature finder . . .	84
5.5	Reconstruction performance on drone imagery	89

Nomenclature

List of Acronyms

ADC	analog to digital converter
AP	average-precision
AUV	autonomous underwater vehicle
BuFF	burst feature finder
CMOS	complementary metal-oxide semiconductor
CCD	charge-coupled device
CNN	convolution neural network
DOAP	descriptors optimized for average-precision
DOF	degree of freedom
DoG	difference of Gaussian
EV	exposure value
FP	false positives
IMU	inertial measurement unit
IR	infrared
LiDAR	light detection and ranging
LiFF	light field feature
LIFT	learned invariant feature transform
mAP	mean average-precision
MMA	mean matching accuracy
NeRF	neural radiance fields
ORB	oriented FAST and rotated BRIEF
PnP	perspective-n-point
RANSAC	random sample consensus
ROC	receiver operating characteristic

R2D2	repeatable and reliable detector and descriptor
SfM	structure-from-motion
SIFT	scale invariant feature transform
SLAM	simultaneous localisation and mapping
SNR	signal-to-noise ratio
SURF	speeded up robust features
TP	true positives

Chapter 1

Introduction

"Curiosity is the essence of human existence"

- Gene Cernan

Curiosity and inquisitiveness have always driven humanity to explore, discover and understand the vastness of the universe. We, as humans, have designed robots to perceive the world beyond the capabilities of our own kind, traditionally to perform strenuous and repetitive work, especially in hazardous environments. As civilization moves toward a new era of technological advancements, robots now, more than ever, have a role to play in building a creative future by perceiving the world in novel ways.

1.1 Motivation

Robots use vision sensors to perceive the environment and perform a wide range of tasks, such as simultaneous localisation and mapping (SLAM) [3, 4], navigation [5, 6], pose estimation [7, 8], depth estimation [9, 10], and 3D reconstruction [11, 12]. A camera, for instance, gives a robot the ability to perceive the world in a similar way to humans, and when combined with other sensors or computational techniques, even beyond human capabilities. By enabling visual sensing, a robot can operate in conditions that are not suitable for humans, such as deep-sea exploration [11, 13],

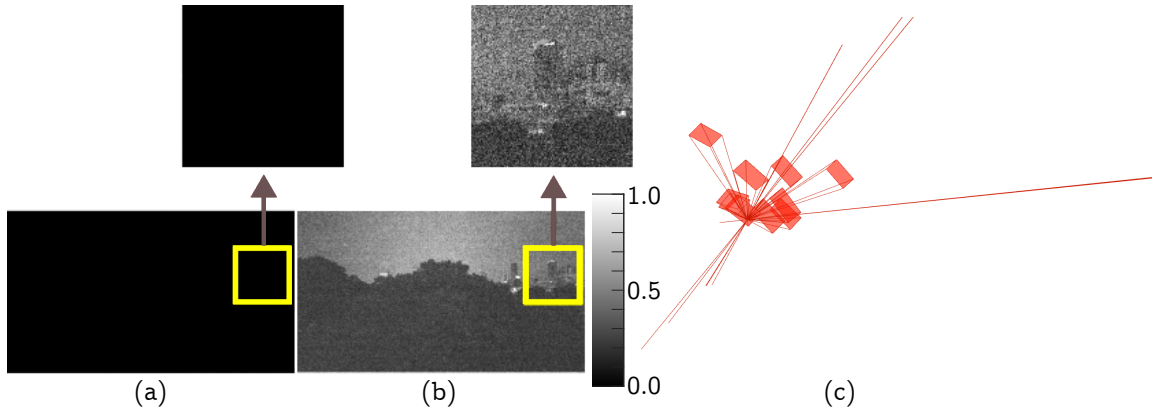


Figure 1.1 – (a) An example of a light-constrained scene captured at night; (b) Contrast-adjusted captured image for visualisation. Reconstructing the scene using 15 images with low signal-to-noise ratio (SNR) images similar to (b) from different viewpoints fails to converge. (c) Relaxing the initializing setting, reconstruction estimates inaccurate camera poses.

mining [14], and space missions [15]. This has significantly accelerated scientific discoveries in recent years [16, 17, 18], and continues to do so remarkably [19].

State-of-the-art vision-based reconstruction methods [20, 21] succeed in good lighting conditions, but they struggle to operate in challenging visual conditions like fog, haze, rain, snow, and low light. We are particularly motivated by the challenge of reconstructing 3D shapes in low light, as current methods yield inaccurate shape estimations, camera trajectories, and can even fail to reconstruct entirely. One such example using COLMAP [20] with scale invariant feature transform (SIFT) feature extraction [22] is shown in Figure 1.1.

Robots may address the issue of low light by using their own light sources to illuminate the environment. However, in some applications, such as studying the behavior of nocturnal animals, it is not possible to use additional light as it can affect their behavior. Smaller platforms might also have weight or power limitations that prevent them from carrying extra light sources. Robots that currently use their own light sources such as an autonomous underwater vehicle (AUV) with strobe lights have an expensive power budget during deployment. So, the use of additional light sources for low-light environments is not always feasible or efficient for robots.

Cameras designed for low-light conditions, such as night vision cameras, use infrared wavelengths to detect dark scenes. Near-infrared (IR) night vision cameras use wavelengths of 700-1000 nm and require their own IR illumination, while those sensitive to 1000-1500 nm use ambient skylight. However, even with ambient skylight, IR illumination is still needed during deployment. Night vision cameras are also expensive [23] and not suitable for mass production of emerging robots [24] such as drone delivery services.

The current approaches in low-light vision for robots mainly focus on improving image quality through denoising [25, 26, 27] or deblurring techniques [28] applied to conventional images. The effectiveness of these approaches is inherently constrained by the limited amount of data present in a single image. While there are computational imaging techniques that can enhance low-light images [29, 30, 31], they often require hardware modifications. In contrast, we present a method that leverages multiple images captured under low-light conditions to maximize the information available for reconstruction and consequently offer a series of guidelines on how to fine-tune camera settings to achieve this goal.

Burst photography is an established mobile photography technique that uses a series of consecutive frames captured over a small exposure time to produce a single image with an improved SNR upon merging [32, 33, 34, 35]. Burst imaging has also been shown to improve the SNR in images without the need for additional ambient light sources [33].

Prior works and ongoing developments on burst imaging are steered towards mobile photography [33, 34, 36, 37], where the primary objective is to produce convincing content for human visual perception while accounting for camera motion due to handshake and scene motion within a single burst of images. On the contrary, achieving satisfactory reconstruction results in low-light conditions necessitates dependable pixel-level matches [38, 20] across various bursts. Given the typical dynamics of robotic platforms during burst capture, which often involve significant occlusion and camera movement between bursts, it becomes crucial to calculate accurate camera pose estimates for reliable reconstruction.

Modern 3D reconstruction techniques generally employ 2D features to find pixel-level correspondences between images [20, 39]. While classical 2D feature extractors [22, 40, 41] and learning-based feature extractors [42, 43, 44] have shown promising results in vision-based reconstruction applications [20], their performance reduces significantly in low-light scenarios [45, 38], making them unsuitable for light-constrained reconstruction. Therefore, there is a need for a new input scheme for reconstruction that can overcome the limitations of classical techniques in low-light environments.

This motivates the development of robotics-driven burst imaging techniques for 3D reconstruction, which could potentially provide reliable pixel-level correspondences and robust camera pose estimates, even in challenging lighting conditions. Such an approach can significantly enhance the accuracy and robustness of the visual perception system for robots and pave the way for new robotic applications in low-light conditions.

1.2 Problem Statement

This thesis seeks to enhance the ability of robots to perform feature-based 3D reconstruction in low-light conditions, which is currently constrained by the low SNR in images captured using conventional monocular cameras. To achieve this, we propose the use of robotic burst imaging for improved feature extraction, which enables robust feature-based 3D reconstruction. The main hypothesis of this research is that burst imaging can provide robots with a better understanding of light-constrained scenes compared to conventional imaging. However, the primary challenge in adopting burst imaging for robotics lies in enabling conventional feature detectors and descriptors to comprehend the information contained within bursts of images to compute reliable features and camera pose estimates. The research aims to explore methods for identifying and utilizing the most useful information within bursts of images, which may involve developing novel feature extractors for robotic bursts to enable light-constrained 3D reconstruction.

1.3 Contributions

The primary focus of this thesis is to bridge the current research gap that exists in enabling robots to operate in low-light conditions using conventional cameras. We establish a new imaging scheme for light-constrained reconstruction by adapting burst imaging for robots to operate at night. We introduce novel feature detectors and descriptors for robotic burst imagery utilizing both physics principles and learning methods. These advancements enable feature-based reconstruction even in light-constrained environments.

The primary contributions of this thesis are:

- We establish the viability of using burst imaging to improve robotic vision in low light ([Chapter 3](#) - [Chapter 5](#)), and provide a set of recommendations for adopting this approach in reconstruction tasks such as structure-from-motion (SfM) ([Chapter 3](#)),
- We introduce a robotic burst feature finder ([BuFF](#)), a 2D + time feature detector and descriptor that finds features with well defined scale and apparent motion within a burst of images for light-constrained 3D reconstruction ([Chapter 4](#)),
- We present a learning framework to find robust light-constrained features from noise-limited bursts of drone imagery captured in millilux condition at night ([Chapter 5](#)), and
- We demonstrate improved feature performance in noise, reconstruction performance including more 3D points, fewer spurious matches and precise camera pose estimates overall using robotic burst imaging compared to conventional methods ([Chapter 3](#) - [Chapter 5](#)).

1.4 Significance

This thesis introduces robotic burst imaging, a novel input scheme for correspondence search for light-constrained 3D reconstruction. By using robotic burst imaging, it is possible to improve the accuracy of correspondence search for reconstruction in low light. This work opens the way for a broad range of applications in which low light commonly complicates vision such as operating autonomous driving and delivery drones at night, mining, and understanding nocturnal behaviors of animals.

By improving the efficiency, and robustness of correspondence search for reconstruction with high-quality features, it becomes possible to create more immersive and realistic virtual environments, as well as to capture and analyze 3D data from remote or inaccessible locations. This also allows more practical and accessible light-constrained reconstruction for a wider range of applications and devices. Overall, by improving the quality of features used in light-constrained reconstruction, this work can help advance the robotics field and enable new applications that were previously challenging or impossible to achieve in low light.

1.5 Outline

[Chapter 2](#) provides the necessary background knowledge to comprehend the thesis as a standalone read. This section covers the fundamental concept of 3D reconstruction, prevalent methods for achieving reconstruction, and various definitions associated with reconstruction applications, specifically [SfM](#). We also include a detailed discussion of existing approaches to light-constrained reconstruction using complementary sensors, as well as the drawbacks of utilizing low [SNR](#) images. We identify the imaging trade-offs with capturing light-constrained scenes using a conventional camera, followed by an exploration of how burst photography addresses these trade-offs. We conclude the section with a discussion on the noise associated with capturing burst imagery in low light.

In [Chapter 3](#), we outline our approach to adapting burst imaging, commonly used in mobile photography for 3D light-constrained reconstruction. We discuss the current trends in burst photography and how all methods, regardless of their intended applications, focus on computing a single higher [SNR](#) image from a single burst of images. We then outline our approach towards robotic burst-based [SfM](#) by using direct methods within a burst of images to improve noise in images and finding sparse feature correspondences between bursts of images to deal with large motions. We evaluate our method against alternative input scheme for 3D reconstruction using conventional images and all images within a burst of images and demonstrate improved performance in terms of feature evaluation, camera pose estimation and convergence rate.

In [Chapter 4](#), we introduce a robotic burst-based feature extractor that leverages apparent motion within a burst of images to avoid misalignment at strong noise. In this chapter, we discuss existing 2D and 3D classical feature extraction methods and identify the need for a robotic burst feature finder. We discuss the typical motion behavior associated with robotic bursts and introduce a burst feature extractor that locates features with well-defined scale and apparent motion via joint search in scale-slope space. We describe these features based on their histograms of edge-orientations. We highlight the overall improved performance in terms of feature evaluation, camera pose estimation and convergence rate against state-of-the-art classical and learning-based feature extraction methods on conventional images and robotic burst-merged images.

In [Chapter 5](#), we develop a novel learning architecture for robotic burst feature extraction to explore light-constrained reconstruction beyond the limitations of classical and physics-based algorithms. We discuss the design choices behind state-of-the-art learning-based feature extraction for reconstruction and identify the research gap in adapting burst imaging for learning-based feature extraction. We design our architecture and loss functions to effectively handle strong noise in low-light imagery while prioritizing the interpretability of burst features, facilitated by confidence maps. Our evaluation against state-of-the-art feature extraction methods on conventional images

and burst imagery shows progressive improvement in reconstruction performance and an ability to operate in lower light conditions than was previously possible.

[Chapter 6](#) concludes the thesis with insights and broader impacts of the research study and highlights future research directions.

Chapter 2

Background

"If I have seen further, it is by standing on the shoulders of giants"

- Issac Newton

2.1 Overview

In this chapter, we provide an overview of background topics and knowledge that support technical development in the later chapters. We first discuss the classes of strategies used in correspondence search in visual reconstruction and analyze existing light-constrained reconstruction methods. We examine a reconstruction example, specifically [SfM](#), and analyze the role of different stages of its pipeline in light-constrained reconstruction. We discuss the existing imaging trade-offs in low-light conditions using conventional cameras in detail. We also introduce burst photography and review its state-of-the-art usage on mobile phones followed by a discussion on noise associated with captured low-light burst imagery.

2.2 3D Reconstruction

During operation, robots equipped with vision sensors, such as cameras, rely on captured images to perceive their surroundings and determine their location. Even though the environment in which robots navigate is three-dimensional, image sensors have a limited design that allows them to capture information in only two dimensions. To accurately localize themselves, robots use multiple 2D images to create a 3D model of the environment. This approach has been effectively employed for reconstruction [20, 39] in well-lit conditions and in various domains, including medical applications [46] and entertainment [47]. Robots perform 3D reconstruction using two classes of strategies:

Direct methods. Direct methods are an effective approach to estimate relative changes in camera poses by utilizing all pixels from input images. Direct methods have proven to be successful [48, 38] in reconstructing images with relatively small motion using monocular cameras [38, 49] and RGB-D sensors that capture both color (RGB) and depth (D) information simultaneously [50, 51, 52].

By considering color, brightness, and intensity gradients, which represent the environment more closely, direct methods provide rich information for robotics applications instead of relying only on particular feature properties [45, 38]. However, direct methods are not suitable for light-constrained reconstruction in scenarios involving large motion, occlusions, and intensity variations, as they assume brightness constancy [38, 45, 50]. This makes an end-to-end reconstruction approach using only direct methods unsuitable [53, 48], particularly for loop closures [54, 38, 45].

Feature-based methods. Feature-based methods are more effective than direct methods in handling large occlusions and motion between images [54, 38, 45]. This is because feature-based methods locate pixels with well-defined features, such as blobs or corners, and describe these features in a way that is invariant to apparent viewpoint and illumination changes [43, 42, 55, 22, 40, 41]. However, state-of-the-art algorithms for feature-based methods struggle under challenging imaging conditions, such as motion blur, low light, and repetitive texture, making it difficult to converge using

visually challenging images [54, 38]. When operating an end-to-end feature-based reconstruction using images captured in low-light conditions, spurious features may be produced during the correspondence search stage, resulting in decreased overall reconstruction performance [38, 54, 53].

Light-Constrained Reconstruction

When capturing images in low-light conditions, achieving high-quality vision-based reconstruction can be challenging because of low SNR in images. One approach that has been utilized to address this challenge involves using complementary sensors in addition to conventional cameras. Recent examples of this include integrating sparse depth measurements from light detection and ranging (LiDAR) with thermal infrared camera for SLAM [56], employing an infrared camera and an inertial measurement unit (IMU) for SLAM [57], and implementing radar-based SLAM [58]. These methods have all shown strong performance in practice. However, to keep emerging robots affordable [59, 60], a low-cost sensing system, e.g., conventional camera, is still the financially advantageous choice, limiting the use of additional sensing modalities within our current scope.

Another approach to dealing with light-constrained images is by leveraging low-light enhancement studies [61, 62, 63, 64, 65] which can directly be applied iteratively over every input image for reconstruction. As most of these low-light enhancement methods are learning-based approaches [61, 62, 63, 65], there is an overhead for collecting datasets that needs to be generalized across different robotic platforms and environments in which robots operate. This is an expensive training requirement for light-constrained reconstruction.

The latest work on domain adaptation [66, 67] and style-transfer methods [44, 68] have shown state-of-the-art performance on day-to-night images, enabling possibilities for more training data for light-constrained reconstruction. Similarly, studies on image synthesis with similar properties to low-light images [69, 70] have recently emerged to design training datasets that will eventually allow robots to operate in low light.

Some studies have taken a step further by acquiring low-light images with autonomous cars [71, 72] as part of comprehensive training datasets highlighting the significance and indispensability of reliable reconstruction in low-light scenarios. While these studies focus on building training datasets for robust reconstruction to overcome some limitations of learning-based approaches in light-constrained reconstruction, we focus on introducing an input scheme that can directly improve correspondence search and reconstruction.

Input scheme for Correspondence Search

Most of the output representations of reconstruction strategies are unique and application-driven though the input scheme is similar and transferable across other

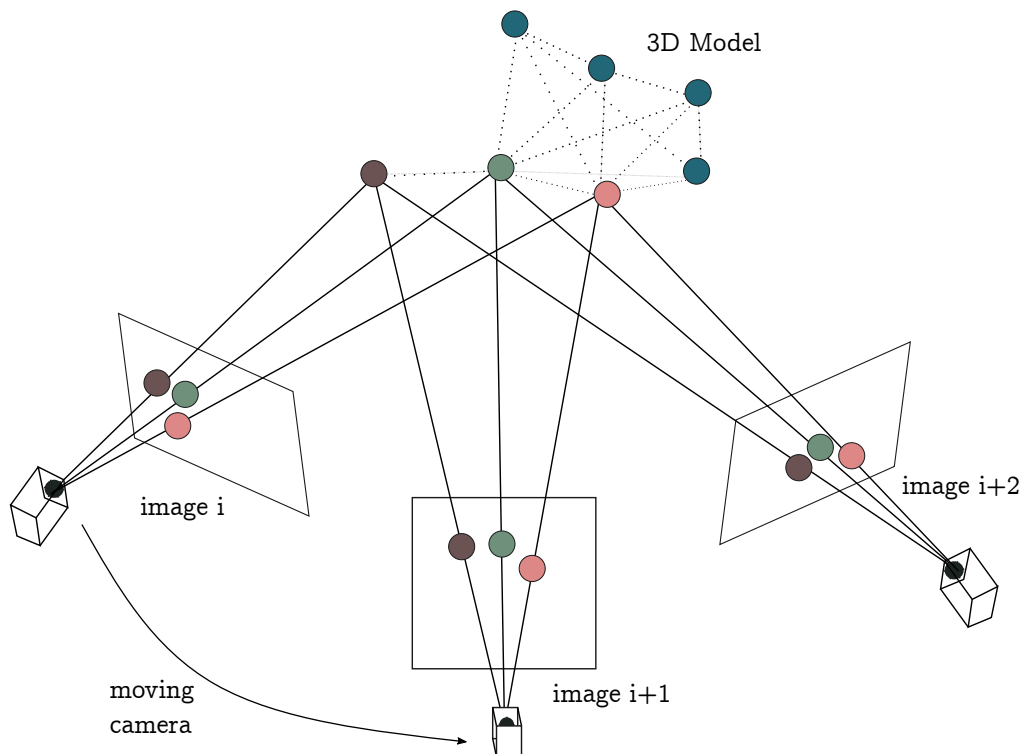


Figure 2.1 – A monocular camera is moved in 3D space and images are captured sequentially. Features are extracted from each of these images, and matched with corresponding features from other images to reconstruct a 3D model. This approach is called structure-from-motion (SfM).

applications. For example, **SLAM** formulates the goal of problem formulation to estimate sensor positions, localize the robot, map the environment, and leverage incremental data using multiple 2D images while **SfM** formulates the outcome as a 3D environment from multiple 2D images captured at different locations [73] as described in Figure 2.1.

While most of the above mentioned light-constrained reconstruction methods are focusing on single-image capture and enhancement, the on-board computation on a robotic platform provides more flexibility in allowing a robot to capture diverse imaging scheme [74, 75, 76]. This also aligns with our observation to capture a novel imaging input scheme to reconstruct a low-light scene.

2.2.1 SfM in Detail

We use COLMAP [20], an end-to-end feature-based state-of-the-art reconstruction pipeline to evaluate light-constrained 3D reconstruction throughout this work. In this section, we aim to provide an overview of the definitions and foundational concepts related to **SfM** as outlined in the COLMAP pipeline in Figure 2.2. By gaining a deeper understanding of these concepts, we can better comprehend the specific design decisions made in subsequent chapters of this thesis. Additionally, we examine the areas of the feature-based **SfM** pipeline that are impacted by the presence of noisy images during the reconstruction process. Through evaluation of **SfM**, we aim to represent 3D reconstruction in general.

Correspondence Search

COLMAP identifies **SIFT** [22] features as keypoints on images and searches for corresponding features across other input images. The correspondence stage of the COLMAP pipeline relies heavily on **SIFT** feature extraction. We discuss feature extraction stages, including detection and description, in greater detail for classical methods in Section 4.2 and learning-based methods in Section 5.2.

Feature Detection. [SIFT](#) locates keypoints, or blob features with well-defined scales, on 2D images by searching a scale space for local extrema. [SIFT](#) is effective at detecting high-quality features on images with high [SNR](#). However, [SIFT](#) performs poorly for feature detection on images captured in low-light conditions [54, 38].

Feature Matching. To provide pixel-level correspondences between images for scene reconstruction, features from multiple images must be matched with each other based on the local appearance of blob features. Compared to other existing classical feature extractors, [SIFT](#) demonstrates stronger matching performance in visually challenging conditions [77, 78]. However, the local appearance of features deteriorates on noise-limited images with low [SNR](#), resulting in poor descriptors for feature matching.

Geometric Verification. Once there are matches across images (putative matches), verifying the matches for geometric consistency (inlier matches) allows robust reconstruction. This is enforced with the epipolar constraint using random sample consensus ([RANSAC](#)) [79]. When using high [SNR](#) images, most of the located features and matches are robust and this stage is used in rejecting outliers based on geometrical constraints. However, in low [SNR](#) images, the verification stage can become computationally expensive due to the overwhelming presence of spurious features and matches to reject [20, 80]. Consequently, it may fail to identify true correspondences, resulting in a less robust correspondence search.

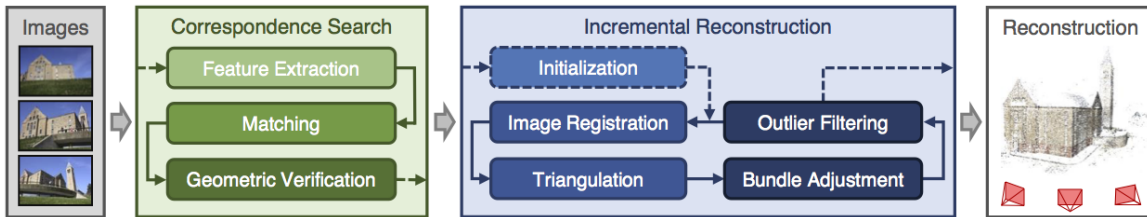


Figure 2.2 – COLMAP pipeline performs incremental [SfM](#) taking sparse input images as the input and producing a 3D reconstruction with corresponding camera pose estimates. The front-end deals with correspondence search including detecting keypoints (features) on the images, finding similar keypoints (matches) between images and geometrically verifying the correspondences. The back-end performs incremental reconstruction with the computed features and matches to register the images for reconstruction. (Source: Schoenberger et al., (2016) [20])

Incremental Reconstruction

By improving the correspondence search stage we expect to improve incremental reconstruction. We evaluate performance metrics associated with this phase of reconstruction across all our methods. The following describes how low SNR images have the potential to demonstrate undesirable results at these stages

Initialization. The initial pair serves as the starting model to which SfM registers new images. SfM continues to register images by using feature correspondences from new images to triangulate points in already registered images (2D-3D correspondences). In this process, SfM addresses the perspective-n-point (PnP) problem of estimating camera poses (i.e., the position of a camera relative to the captured scene or 3D structure) and camera intrinsics (i.e., internal parameters of a camera such as focal length and radial distortion) [81]. However, low SNR images tend to produce more spurious features. This can lead to unreliable estimates at this stage, potentially causing a failure to reconstruct the initial pair and, eventually, a failure of the overall 3D reconstruction.

Image Registration. Once initialized, the incremental stage involves registering every input image through triangulation, bundle adjustment, and outlier rejection, as shown in Figure 2.2. Since incremental SfM has to deal with large occlusions and camera motion between each image, the estimates heavily depend on the robustness of the correspondences. This robustness might be compromised when input images have a low SNR. Consequently, this can hinder the use of input images, leading SfM to disregard them entirely.

Triangulation. It is possible to determine the locations of a 3D point as seen from multiple cameras using verified feature locations and known camera positions. This is the converse problem to camera pose estimation. In SfM, a new scene point is triangulated and added to the 3D model when at least another image from a different viewpoint is registered. This increases the stability of the model. However, feature-based methods are prone to a large number of outliers due to inaccurate two-view verification of putative matches along the epipolar line. In general, a single mismatch

can be as costly as merging two or more independent points. Low-light images can further amplify this error.

Bundle Adjustment. Inaccuracies in either image registration or camera pose estimation increase triangulation error when computing 3D points. Non-linear refinements of camera parameters [81] and point parameters using Levenberg-Marquardt [82] minimise reprojection error and improve reconstruction using high SNR images. However, images captured in low-light conditions have relatively low reliable estimates for bundle adjustment.

2.3 Imaging Trade-offs

In the previous sections of this chapter, we recap our motivation to allow robots to operate in light-constrained scenes using a conventional camera and discuss existing alternative light-constrained reconstruction methods. We also establish the background knowledge required to understand a reconstruction method, e.g., SfM and key factors, e.g., inlier matches, pose estimates, number of registered images, and number of 3D points, to look out for that corresponds to different stages of reconstruction while designing our approach for SfM. In the following section, we discuss why it is difficult to capture images at night by exploring the imaging trade-offs of a camera and how we can leverage the trade-offs to introduce an input scheme for reconstruction.

For a given viewpoint and lighting condition, a robot considers any image as a point lying in the 3D space defined by the controls of the camera parameters: exposure time, aperture settings, and analog gain settings. This is because when a camera captures a scene it detects the number of photons arriving at each pixel on the sensor, and the number of photons varies based on exposure time – how long the sensor is exposed to light, aperture – the opening diameter through which light enters, and analog gain – the amplification applied on the captured signal before digitalization.

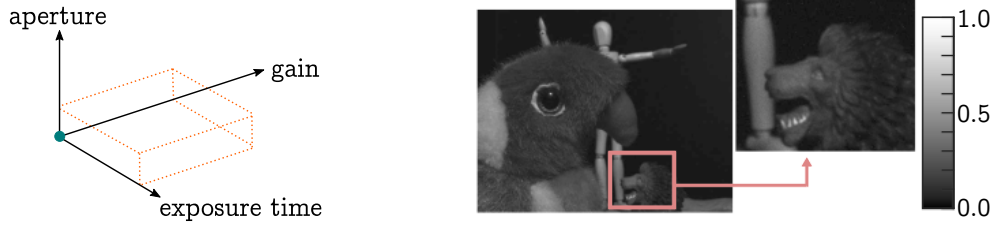


Figure 2.3 – Capturing an image in a well-lit condition (80 lux) with minimum gain and minimum exposure time with narrow aperture for high SNR sharp all-in-focus image. (Contrast-adjusted image for visualisation)

Images captured in well-lit conditions have high SNR as in Figure 2.3. Existing state-of-the-art datasets in day light and gold-standard images for robotic vision applications are captured with similar setting [83, 84, 85]. Capturing an image with the same settings at night generates an extremely dark image, which doesn't represent the environment with enough visual information as in Figure 2.4. Images captured at night have low signal and so, a robot that uses such images to reconstruct the environment it navigates struggle to localise itself.

Aperture. The opening diameter of an aperture determines the amount of light that enters the sensor. A wider aperture allows more light to enter, resulting in a high signal up to the saturation limit. However, widening the aperture reduces the depth of field, as demonstrated in Figure 2.5. Depth of field refers to the range of distances between the nearest and farthest objects that appear acceptably sharp in an image. Because we operate feature matching throughout the image, the background of the image is as important as the foreground of the image. Therefore, the maximum aperture is limited by its impact on depth of field in feature-based reconstruction.

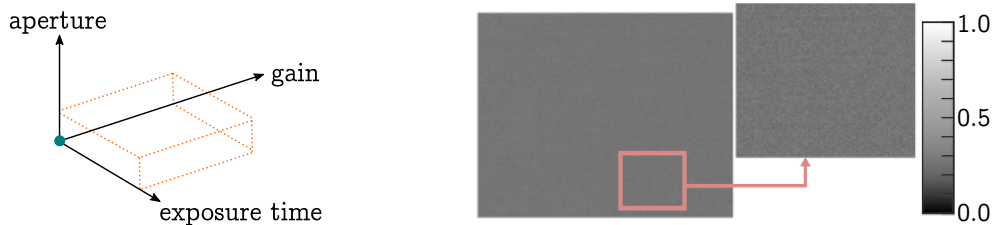


Figure 2.4 – Capturing an image at night with similar setting as well-lit conditions results in extremely dark image because there is limited signal. (Contrast-adjusted image for visualisation)



Figure 2.5 – Capturing an image at night with wide aperture while keeping the other camera parameters minimum allows more signal, but results in narrow depth of field. (Contrast-adjusted image for visualisation)

Exposure. The duration of the shutter being open, known as exposure, determines the amount of light that enters the sensor. Longer exposure times allow more light to reach the sensor, producing a higher signal until the saturation limit is reached. Unlike traditional photography, which involves capturing images from a stationary point such as by using a tripod, robots usually capture images while in motion. This can result in motion blur in the captured images, as shown in [Figure 2.6](#).

There are also dynamic objects in the environment that can cause motion blur, resulting in poor reconstruction. However, [SfM](#) design ignores such artifact during reconstruction. Our motivation is to give independence to the robot in terms of navigation which encourages us to capture images by keeping the exposure as long as possible without causing unacceptable levels of motion blur for robust reconstruction.

Analog gain. When the signal is low either because the environment is light-constrained, or the aperture is narrow, or the exposure time is short, increasing the gain of the amplifier is beneficial. Increasing the analog gain amplifies both the signal



Figure 2.6 – Capturing an image at night with long exposure while keeping the other camera parameters minimum allows more signal, but introduces motion blur. (Contrast-adjusted image for visualisation)

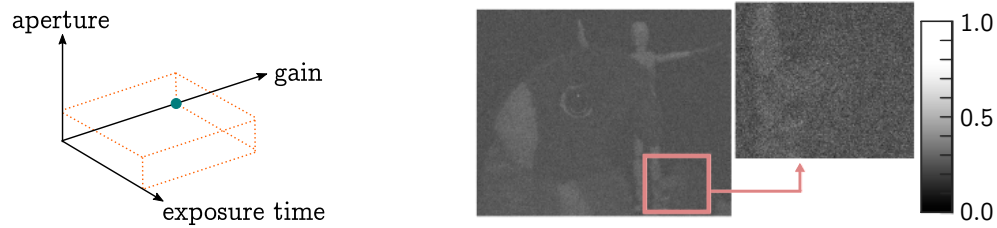


Figure 2.7 – Capturing an image at night with high gain while keeping the other camera parameters minimum amplifies the signal and noise. (Contrast-adjusted image for visualisation)

and the noise as in [Figure 2.7](#), allowing analog to digital converter ([ADC](#)) to operate above the minimum quantization level.

Saturation. For a captured image to be useful for reconstruction, it is important to select an overall setting that is not saturated at the higher bound with excessive light or lower bound with barely any light. This allows us to select a gain that can be applied before losing information in bright parts of an image. This is unavoidable while capturing some scenes, e.g., street light captured at extremely low light, and needs to be treated correctly.

2.4 Burst Imaging

Capturing images that have large overlap among each other over a small exposure is called *a burst* photography. Burst photography is a technique traditionally used in cameras, and eventually on mobile phones to capture sport motions or motion in general which are unpredictable as shown in [Figure 2.8](#). Burst photography is widely utilized in sports photography [[86](#), [87](#)], enabling users to choose the best image from

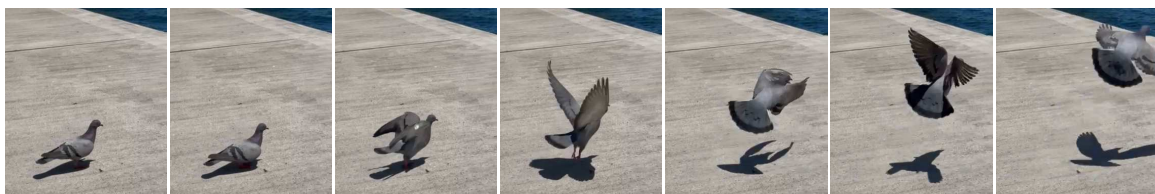


Figure 2.8 – Traditionally, burst photography used to capture unpredictable motion to select the best frame for visualisation e.g., a bird flying.



Figure 2.9 – (left) Burst photography captures all-in-focus images over small exposure time with high gain to overcome imaging trade-offs. (right) An example of a single burst of images: multiple images with high overlap and small exposure.

a sequence of consecutive frames, such as capturing a basketball going through the hoop or a football goal.

Capturing multiple low [SNR](#) images from a static camera as shown in [Figure 2.9](#) and temporally merging them improves the signal in an image, according to [\[73\]](#). However, capturing multiple images from a static platform in robotics requires a high budget for deployment or accurate image stabilization as small variations in motion can result in large pixel differences among images.

Google recently introduced burst imaging [\[33\]](#), a computational method that captures multiple underexposed images with minimal motion variation. This leverages burst photography techniques to capture multiple images. But, instead of visualizing motion within a burst to select a single desired scene, burst imaging aligns each frame within a burst to a common image (i.e., the initial frame of the burst) and merges them temporally to obtain a higher [SNR](#) image, enabling low-light photography. Burst imaging yields high-quality images for night photography on Pixel mobile phones [\[33, 34, 35\]](#). This technique involves capturing 15-30 frames in a single burst and subsequently merging them to produce a single image with minimal noise levels and motion inconsistencies, resulting in visually appealing images for the human visual system.

2.5 Noise Modeling

Most of the state-of-the-art robotic vision algorithms used in reconstruction [22, 42] are noise tolerant to a certain degree by performing outlier rejection to identify robust features and computing inlier matches for reconstruction [39, 20] as discussed in previous section Section 2.2.1. As we are motivated to improve 3D reconstruction by improving correspondence search to identify high quality features, we discuss the noise associated with captured imagery in the following.

A camera captures a scene in the presence of light, or more precisely by detecting the number of randomly generated events – a discrete number of photons arriving on the sensor. There is a broad range of noise associated with captured imagery and most of them originate from camera electronics. One source of noise that purely originates from the scene is Photon noise.

Poisson noise. Also known as photon noise, the dominant noise on a high SNR image is Poisson noise [88]. This is independent of camera electronics and occurs purely based on the captured scene and demonstrates a Poisson distribution of discrete

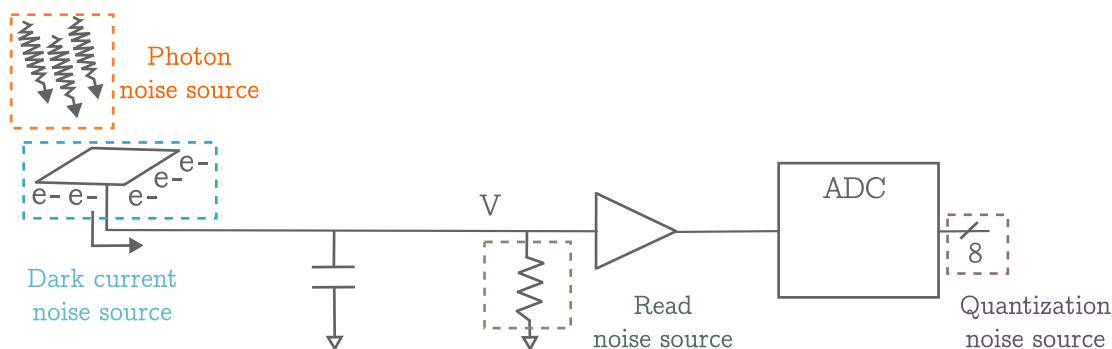


Figure 2.10 – Considering a simplified model of a camera, the photoelectric effect converts a set of photons arriving at a single pixel to electrons. Once the capacitor charges using these electrons, it discharges electrons through a resistor to allow amplification of the analog signal. The amplified analog signal then goes through ADC to produce a digital signal. Noise occurs at different stages of the flow - Poisson noise (orange) depends on the scene, dark current noise (blue) relates to thermally generated electrons at pixels, read noise (green) occurs before the amplifier at resistors – the dominant source of noise in our captured imagery. Quantisation noise (violet) is the noise associated with converting analog signal to digital number. (Source: adapted from Dansereau (2021))

events. In this thesis, as we deal with low SNR images, our dominant source of noise comes from the camera electronics.

Dark-current noise. When a set of photons arrives at the sensor, based on the quantum efficiency of the sensor, photons are converted to electrons. This conversion occurs due to the photoelectric effect – release of electrons when light hits the surface. Dark current noise occurs at this stage corresponding to the thermal activities of the electrons. Dark current noise is negligible when images are captured with an exposure less than a second in general. For images captured over longer exposure, this noise is modeled as a Gaussian distribution.

Read Noise. Read noise is a combination of system noise components. The on-chip output amplifier sequentially transforms the charge collected at each site into a measurable voltage. The amplifier generates zero mean read noise that is independent of the number of collected electrons [89]. This is unavoidable during imaging and occurs mainly in the terminating resistor just before the amplifier. This noise is uniformly added to all pixels in an image and is also modeled as a Gaussian distribution. Read noise is the dominant source of noise in low-light imaging.

Quantization Noise. The captured signal is analog and doesn't use any conversion for film-based cameras. However for digital imaging with charge-coupled device (CCD) and complementary metal-oxide semiconductor (CMOS) sensors, analog signals are converted to digital numbers. When an analog signal is digitized, it is divided into a finite number of discrete levels or quantization levels, resulting in a loss of information. Quantization noise refers to the difference between the actual analog signal and the quantized digital signal [90]. The level of quantization noise depends on the number of bits used to represent the analog signal. The more bits used, the finer the quantization levels and the lower the quantization noise. Conversely, the fewer bits used, the coarser the quantization levels and the higher the quantization noise. Images captured in low light have low signal and in such images quantization noise is strongly noticeable [91].

Fixed pattern noise. Sensor pixels have variations among each other due to the silicon manufacturing process. These are specific to cameras and vary between cam-

eras of the same models. This represents fixed locations of darker and brighter pixels than most of the pixels on the sensor and does not show significant variation over time [92, 93].

Signal-to-noise ratio. The signal to noise ratio of an image is defined as in

$$SNR = 10 \log\left(\frac{\sigma_{signal}^2}{\sigma_{noise}^2}\right), \quad (2.1)$$

where σ_{signal} represents the amplitude of the signal, i.e., the standard deviation and σ_{noise} represents the amplitude of noise in the signal. An image captured in low-light conditions demonstrates strong noise because it has a low signal and noise dominates the captured image. Note, all images have noise.

Chapter 3

Light-Constrained Burst SfM

*"Wonder is the feeling of a philosopher, and
philosophy begins in wonder"*
- Socrates

3.1 Overview

We set up the required background knowledge to understand the following technical chapters in [Chapter 2](#) and establish an understanding of noise models associated with captured imagery. In this chapter, we present a robotic burst-based structure-from-motion, an imaging computation method for 3D reconstruction in low light, by adapting burst techniques in the photography domain. We define burst imaging for robots, similar to [\[33\]](#), as capturing multiple frames in rapid succession over a short exposure time.

To recover images from which features can more reliably be extracted for feature-based reconstruction, we adapt burst imaging to reconstruct 3D scenes in low light. We do this by combining feature-based [SfM](#) and burst photography, exploiting the advantages of direct methods for image registration [\[38\]](#) and applying feature-based [SfM](#) [\[20\]](#) to handle large camera motions between robotic bursts.

Our approach assumes that the level of noise present in images is not so great as to make information unrecoverable. This occurs in extremely low-light images as described in [Chapter 2](#) when the underlying signal is suppressed by quantization noise. We also assume that there exists a sufficient overlap between consecutive images within robotic bursts for registration, which is commonly available for mobile platforms moving at moderate speeds relative to their environment. This work opens the way for a broad range of applications in which low light commonly complicates vision such as autonomous driving and delivery drones at night.

To validate our method, we mounted a monocular machine vision camera on a UR5e robotic arm as in [Figure 3.1](#), and captured burst imagery of real scenes in an illumination-controlled environment over different exposure times to demonstrate strong and moderate noise on burst imagery. We demonstrate improved feature performance, reconstruction performance and camera pose estimates against alternative approaches.

This work is a first step toward solving the problem of 3D reconstruction in low light using burst imagery. Parts of this chapter are published as [\[1\]](#) and the code and dataset are available at: <https://roboticimaging.org/Projects/BurstSfM/>.

In summary, our key technical contributions are:

- We establish the viability of using burst imaging to improve robotic vision in low light and provide a set of recommendations for adopting this approach in reconstruction tasks such as [SfM](#),
- We evaluate different approaches to burst imaging in robotics applications and show that burst capture with merge offers significant advantages in both computational requirements and performance; we offer a theoretical explanation and experimental evidence showing why this is so, and
- We demonstrate the proposed method improving low-light [SfM](#) by yielding more precise 3D points, higher quality features, fewer spurious features, more precise camera trajectories, and an ability to operate in lower light than was previously possible.

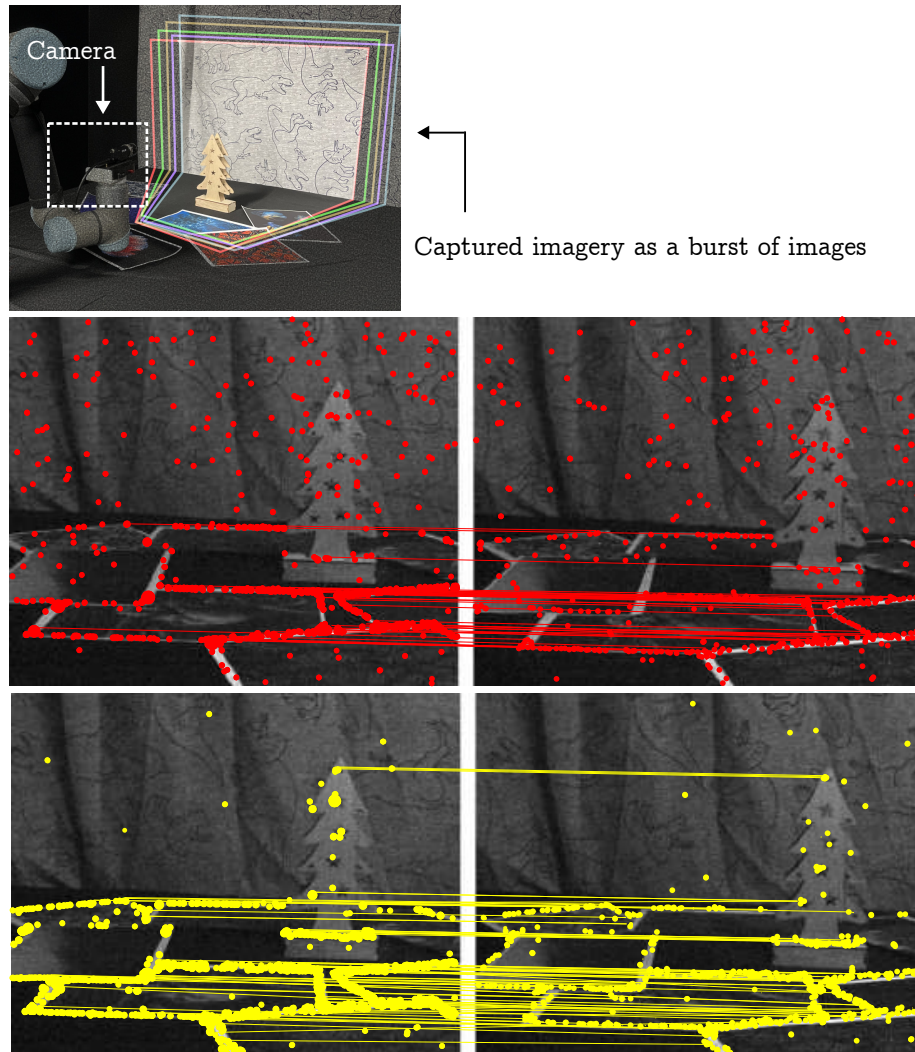


Figure 3.1 – Visualisation of light-constrained scene. (top) Robot arm-mounted machine vision camera; (middle): red: 43 matching pairs between consecutive conventional noisy images which fail to converge for reconstruction; (bottom) yellow: 110 matching pairs using our proposed merging approach. It is important to note that 36% of input images of this particular scene fail to find any matches in the conventional noisy approach where as all input images are registered using our method for reconstruction. Our approach provides a more stable set of high-quality true features and less spurious features for motion-dependent applications such as [SfM](#).

3.2 Related Work

Alternative exposure schemes. When optimizing camera exposure settings for low [SNR](#) images, there exist many trade-offs in the final image quality as discussed in [Section 2.3](#). Considering alternative exposure schemes, depicted in [Figure 3.2](#),

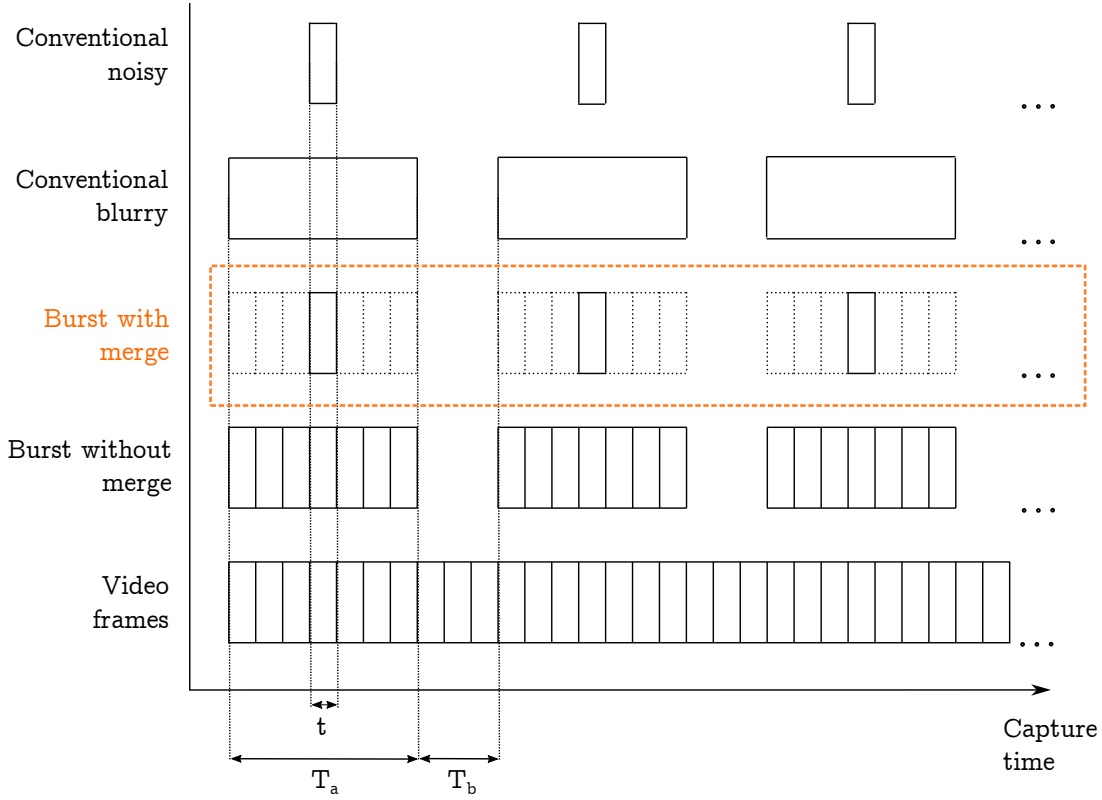


Figure 3.2 – Alternative exposure schemes for capturing images for light-constrained 3D reconstruction. T_a is the duration of an N -images robotic burst, t is the duration of a single frame, and T_b is the delay between robotic bursts.

increasing the exposure duration for each frame gathers more light at the expense of increased motion blur and loss of information. Capturing images as a sequence over a long period of time yields a video. However, it is computationally expensive to process images at full video rates and in low light.

Single-image based approaches. Conventional noisy images can be pre-processed with denoising approaches [25, 94] to get high SNR. However, these approaches are fundamentally limited by the amount of information in a conventional single image. They also have a tendency to produce visually pleasing results and not necessarily accurate information through learned image priors [37]. This affects the reliability of estimates for applications such as reconstruction and tracking.

Learning-based methods. Learning-based techniques show promising results for denoising [95, 96, 97], deblurring [98, 99, 100], dehazing [101, 102, 103] and image

enhancement [104, 105, 106]. However learning-based approaches make no guarantee of generalising beyond their training domains [107, 108, 109] and compared with the proposed non-learning-based approach, these require collection of appropriately scaled datasets and corresponding training times. Learning-based approaches also present limitations around interpretability and direct understanding of the structure by which a model works [110, 111].

Burst photography. Utilizing the advantage of time slices as described in [112], burst photography has emerged as a promising technique for night photography. This approach involves capturing multiple images over a short time period and then merging them to generate a single image with high SNR as discussed in Chapter 2. The concept has been further developed and implemented in various forms, such as the night sight mode in Google Pixel phones [33], astrophotography mode [34], and multi-frame super resolution [35], as well as using learning approaches [113]. Recently, the concept has been extended to quanta burst photography [114], which uses single-photon sensors to produce high-quality images.

Previous and state-of-the-art work in burst photography [32, 33, 34, 35] has mainly focused on generating visually appealing still images from bursts, accounting for hand-shake motion that is typically different from robotic platform dynamics and low-light conditions in which robots operate. Our work considers more challenging imagery than what is typically employed in existing burst photography work [33, 34]. Another key difference to photography is that robotics typically covers extended areas over long trajectories, meaning multiple bursts of images become necessary.

In this work, we adapt burst imaging to reconstruct 3D scenes in low light. We do this by combining feature-based SfM and burst photography, exploiting the advantages of each. By capturing a rapid successions of frames, we enable the use of direct methods for image registration [38], exploiting the small camera motion between images to yield a strong SNR advantage with moderate computational expense. After extracting features from the merged burst images, we apply feature-based SfM [20] to handle large camera motions between bursts. Feature-based methods handle large camera

motions [39] while benefiting from the improved feature quality associated with the robotic burst-merged images.

Because our method compresses each robotic burst into a single denoised image, the total computation during reconstruction is lower compared to an approach that employs all measured frames as described in Figure 3.2. Skipping the merge step directly ingests all images in the robotic bursts into the SfM pipeline and we show this to be less effective than merging in terms of both performance and computational expense. Furthermore, the higher SNR in our imagery can reduce the numbers of spurious features, again lowering computational requirements during reconstruction as there are fewer outliers to detect and reject.

3.3 Robotic Burst-Based Structure-from-Motion

In this section, we consider the salient aspects of burst photography [33], and explain its integration into an SfM pipeline. As in [33], we capture multiple images, establish dense correspondence between them, then merge the aligned stack with a temporal voting scheme as shown in Figure 3.3. Taking this direct approach within the robotic burst exploits the temporal coherence between frames, manifesting as relatively little illumination variation and occlusion/disocclusion within a robotic burst. We then extract features from each robotic burst-merged image to deal with appearance changes associated with the larger timescale and translation between robotic bursts. In this work we further introduce and evaluate spatial Wiener and bilateral filtering.

The following outlines image acquisition, alignment, merging, and integration into a reconstruction pipeline. Further practical considerations are discussed in Sec. 3.5.

3.3.1 Image Acquisition for Reconstruction

As depicted in Figure 3.2, we capture multiple robotic bursts over a trajectory, with each burst containing N frames and each frame having exposure time t , taking time T_a per robotic burst with a delay T_b between robotic bursts.

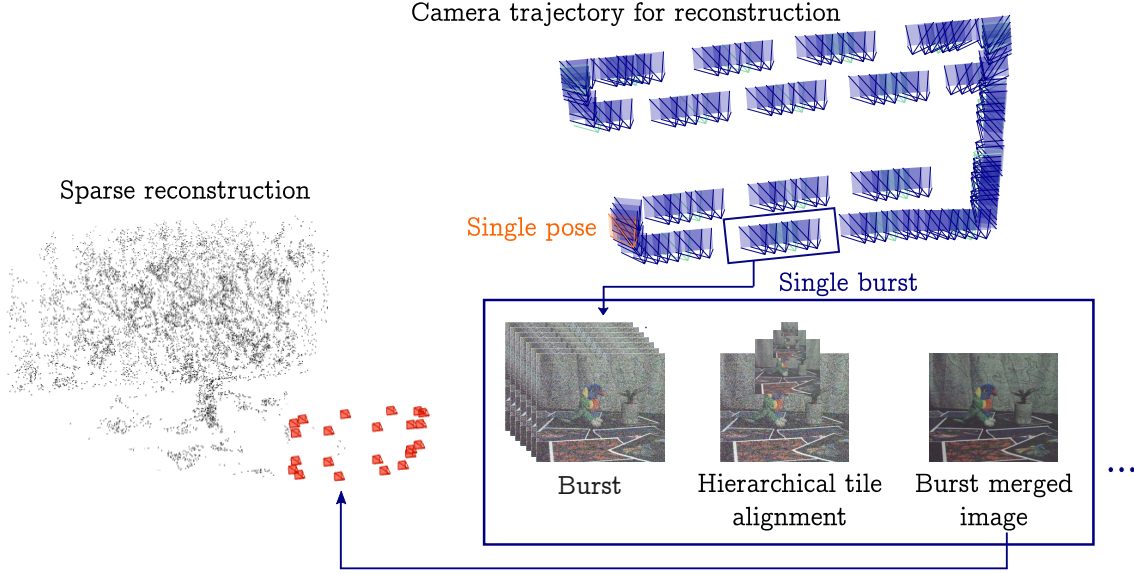


Figure 3.3 – We capture multiple robotic bursts as shown in camera trajectory for reconstruction. As we capture the images on the fly, each robotic burst has multiple images with high overlap and small motion between them. We employ direct methods within each robotic burst to handle noise and feature-based methods between robotic bursts to handle large camera motion. We use temporally merged images of each robotic burst for sparse 3D reconstruction.

Capturing a greater number of images per robotic burst increases the resulting SNR in merged images, provided there is a sufficient image overlap. However, more images are also more computationally expensive for capturing, buffering and processing, so there is a trade-off in selecting the number of images in a robotic burst between quality and computation. This is application-dependent: in our experiments, we use robotic bursts of $N=7$ frames based on empirical evaluation. – see further discussion in [Section 3.5](#).

3.3.2 Robotic Burst Alignment

To address non-uniform apparent motion between the frames in a burst, we perform hierarchical tile alignment, a coarse-to-fine alignment on multi-level Gaussian pyramids of single-channel images similar to [33]. We use a patch of pixels as a tile and using the initial estimates from the coarser scale, we compute pairwise-tile-based

alignment at each pyramid level by minimizing the distance measure. This is the absolute residuals between the common tile and the corresponding candidate tile of each alternate frame within a robotic burst [73]. We perform sub-pixel alignment at coarse scale to improve initial estimates and pixel level alignment by minimizing L1 residuals [115] at the fine scale of the pyramid.

An example of a common image and an alternative image and the corresponding motion map between the images at each level of the hierarchical pyramid is shown in Figure 3.4. Note, while mobile photography uses initial frames of a burst as common image, we merge images in a burst ideally to the middle frame for robotic applications, to address for the relative motion between frames, improving the information over an intermediate time. We empirically tune the parameters of the alignment process to strike a balance between processing time and tolerance to motion between the frames. Practical guidance for tuning alignment parameters is provided in Section 3.5. By

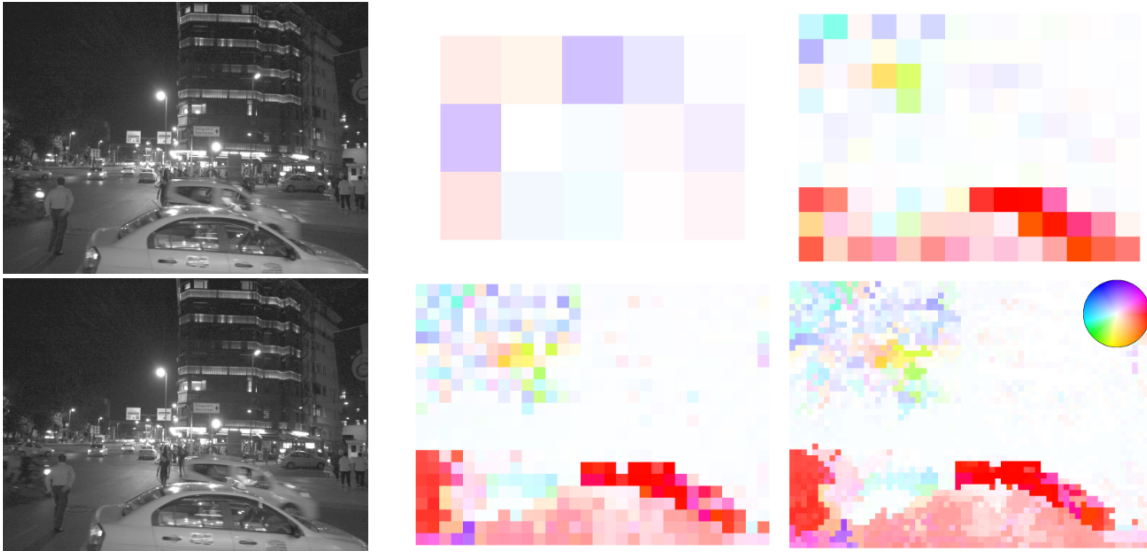


Figure 3.4 – Intermediate results on hierarchical tile-based alignment between common image (top left) and an alternative image (bottom left). Four scales from coarse (top middle) to fine (bottom right) is shown here which demonstrates various apparent motion within a robotic burst. This also shows objects closer moving with larger apparent motion i.e., red tiles representing the foreground car than the far away objects in the scene i.e., lighter pixels representing building at the background. (Source: Hasinoff et al., (2016) [33])

aligning each image within a robotic burst to the common image of the robotic burst, we have an aligned stack of images that corresponds to the common image.

3.3.3 Robust Temporal Merge

We merge all the aligned images within the aligned stack following the pairwise method presented in [33]. To increase robustness to motion, this employs temporal filtering with a per-image contribution in the frequency domain as in

$$\overline{T}_R(\omega) = \frac{1}{N} \sum_{z=1}^N T_z(\omega) + A_z(\omega)[T_R(\omega) - T_z(\omega)], \quad (3.1)$$

where N is the number of frames within a robotic burst, $T_R(\omega)$ is common frame, $T_z(\omega)$ is frame to be added to the estimate, and $\overline{T}_R(\omega)$ is the updated estimate.

A_z is the contribution of each frame, found as in

$$A_z(\omega) = \frac{|D_z(\omega)|^2}{|D_z(\omega)|^2 + c\sigma^2}, \quad (3.2)$$

where $D_z(\omega) = T_R(\omega) - T_z(\omega)$, σ^2 is the noise variance and c is the degree of contribution that increases noise reduction at the expense of robustness to misalignment. Robotic burst imaging does not degrade image quality for fast motion but the quality gain decreases as overlap between images decreases.

Misaligned information is suppressed with the voting scheme as they present a large difference D_z between the common and alternative frames, placing a higher weight on the common frame. Ringing artifacts that are commonly associated with frequency-domain filters [116] are avoided with a raised cosine windowing approach – see [33] for further detail. We also evaluate addition of a noise shaping Wiener filter and an edge-preserving bilateral filter for further spatial noise reduction as in Figure 3.5.

Figure 3.6 illustrates typical performance of the align-and-merge approach, compared with conventional capture which yields noisy imagery, and naïve averaging which is sensitive to motion. Burst with merge shows improved SNR, ideally improving

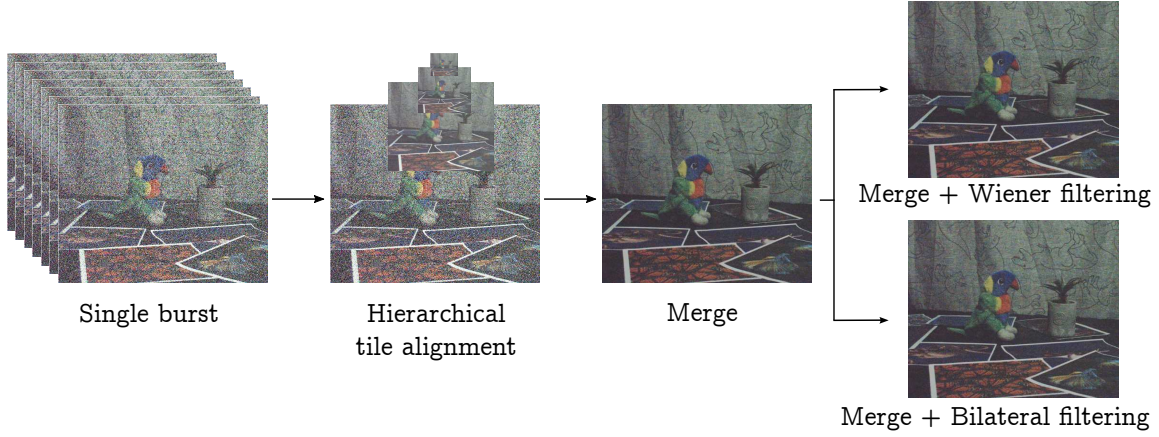


Figure 3.5 – We improve the noise in the captured images by aligning every image in a burst with the chosen common image of the burst (i.e., middle frame for a robotic burst). We merge the aligned images in temporal direction with a voting scheme to avoid misalignment. We use Wiener and bilateral filtering on temporally merged images. We give the outputs of this pipeline to COLMAP as three different inputs: burst with merge, burst with merge and Wiener filtering and burst with merge and bilateral filtering for reconstruction.

SNR, the ratio between the power of signal and noise by $\sqrt{N}$ for N -image robotic bursts [112].

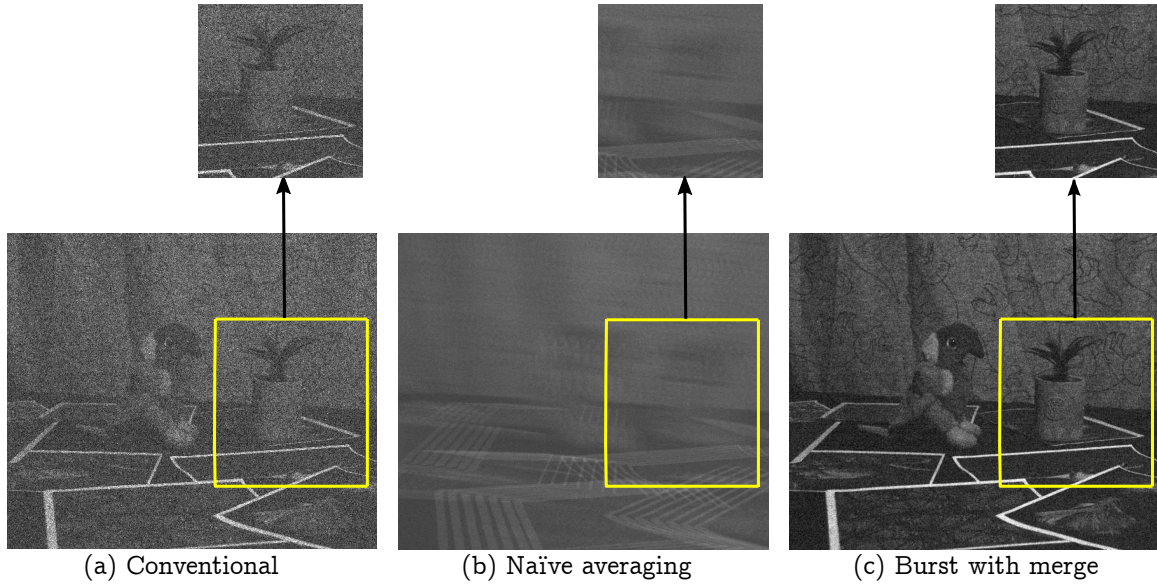


Figure 3.6 – Application of alternative exposure schemes: (a) Single frame captured over an exposure time of t ms. (b) Multiple frames captured over T_a ms and merged naïvely – this is representative of a long exposure scheme with motion blur. (c) The proposed merging approach yields a clearer image without blur.

Reconstruction Pipeline

We use COLMAP [20], an end-to-end feature-based state-of-the-art reconstruction pipeline to extract, match and geometrically verify features between sparse camera poses. We demonstrate sparse reconstruction by triangulating scene points and refining via bundle adjustment. We evaluate performance following reconstruction and feature performance metrics laid out in [20, 117]. For background understanding, please refer to Chapter 2.

3.4 Results

In the following we first evaluate improvements in feature detection and accuracy afforded by our method in noisy imagery using synthetic image sequences. By employing synthetic scenes we offer greater control over noise and scene content similar to prior quantitative feature evaluations [117, 118]. Then we quantitatively evaluate our method in an SfM pipeline, comparing against conventional image capture and direct use of all images in the robotic burst. We consider both 3D reconstruction performance and camera trajectory accuracy.

3.4.1 Feature Performance in Noise

We generate a synthetic set of images with known feature locations to demonstrate feature performance in different noise levels. We use synthetic noise which correspond closely to the noise levels of the captured imagery appearing in the following sections.

The synthetic images, shown in Figure 3.7, have 25 disks at varying scales. The disks have an apparent motion of at most 12 pixels per frame. The contrast between the disks and background is 51:26 where 255 is white. We normalise the intensity of the images between 0 and 1 and introduce moderate noise with variance 0.03 (top) and strong noise with variance 0.1 (bottom).

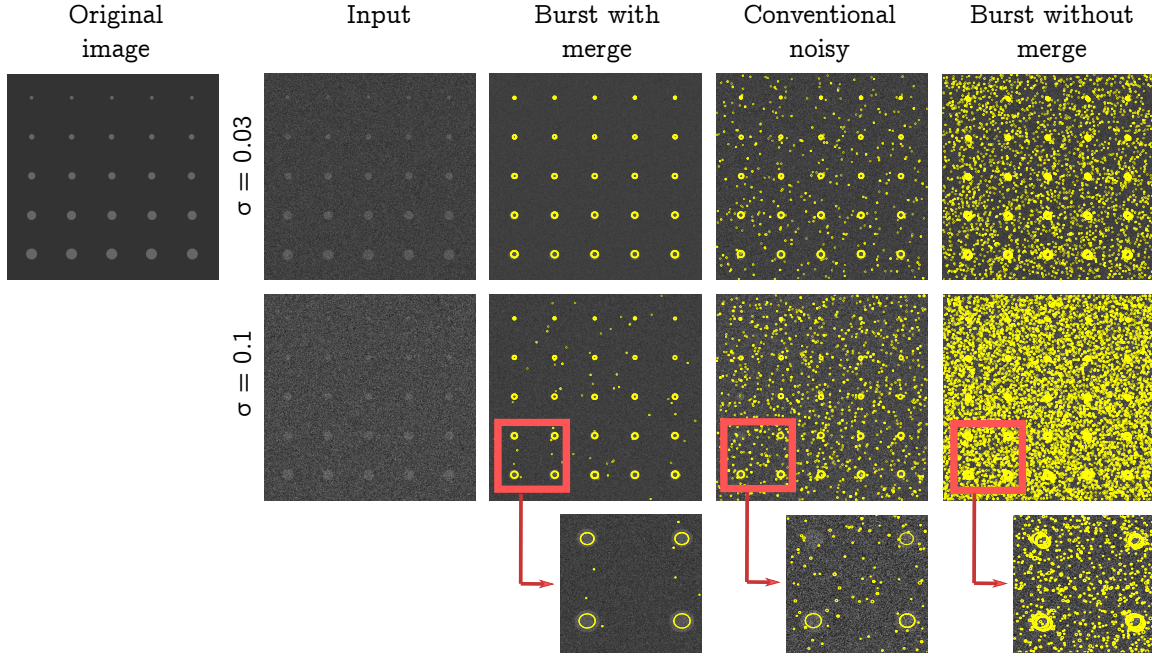


Figure 3.7 – Detected features on synthetic burst of images using [SIFT](#), original common image (top row, left) - presented at two noise levels σ . At the lower noise level (top row), our method performs well with no spurious features, while conventional noisy method and burst without merge generates more spurious features. At strong noise (bottom row), while our method detects all true features, conventional noisy approach detects more spurious features and fewer true features than ours and burst without merge registers overwhelming spurious features.

We extract [SIFT](#) [22] features at a peak threshold of 0.015 on the synthetic image burst with known feature locations. It is evident from [Figure 3.7](#) that burst with merge performs better than burst without merge and conventional noisy on corresponding common image with no spurious features for the moderate noise case ($\sigma = 0.03$). While our method extracts all true features in strong noise, the conventional method fails to extract all true features and burst without merge is overwhelmed by spurious features.

We demonstrate the measurement of true positives ([TP](#)) rate - the number of keypoints detected that corresponds to the ground truth feature locations of the synthetic common image and false positives ([FP](#)) count - the number of spurious features detected quantitatively. The top row shows the [TP](#) rate and [FP](#) count at peak detection thresholds of 0.006 and 0.01 for a sweeping noise level from 10^{-7} to 10^1 in synthetic

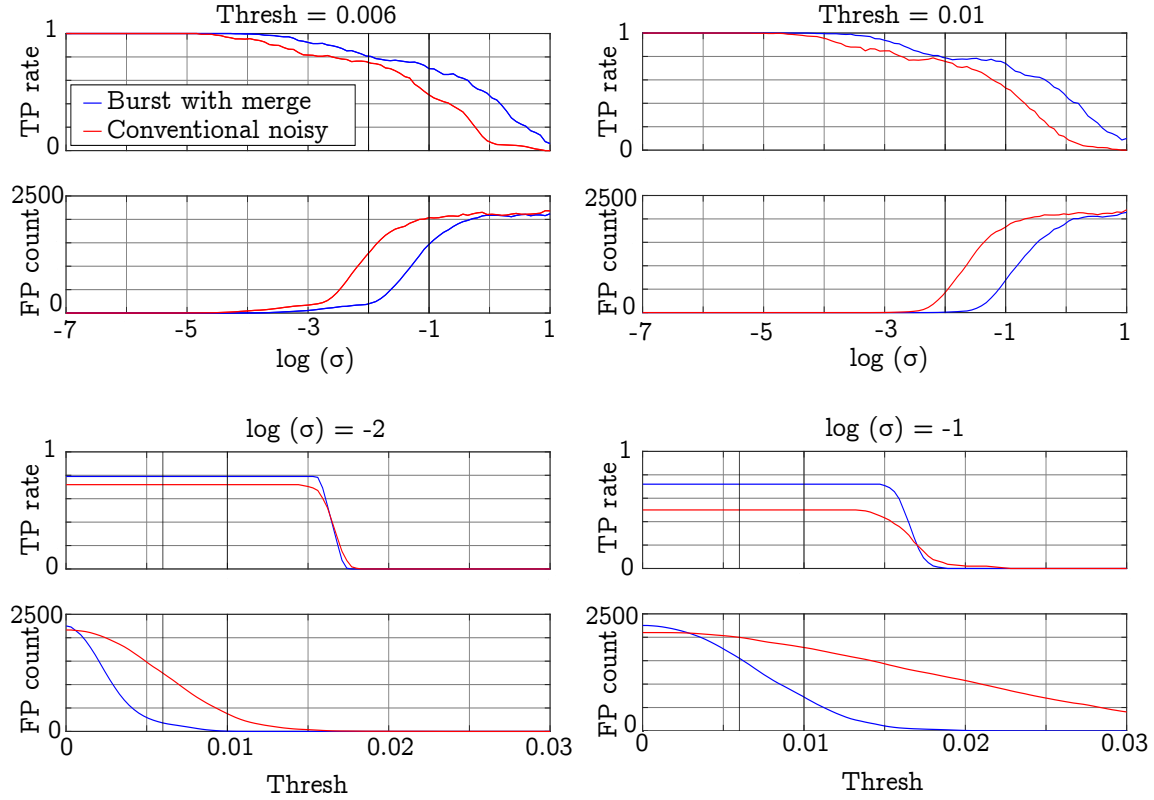


Figure 3.8 – Noise performance: (top row) Sweeping noise level σ for fixed detection thresholds, our method (blue) shows a higher true positive (TP) rate and lower false positive (FP) count than conventional imaging (red). (bottom row) Sweeping detection threshold, our method delivers a much higher TP rate in high noise and lower FP count for appropriately set threshold than conventional imaging. Overall, our method matches or outperforms conventional method in feature performance in noise.

burst imagery and the bottom row shows the TP rate and FP count for moderate and strong noise levels for a sweeping peak threshold level (varying peak threshold detection values) from 0 to 0.03 as shown in Figure 3.8.

As noise increases from left to right that correspond to two specific peak thresholds, we observe an increase in spurious features and decrease in true features. However, our method consistently outperforms conventional noisy capture. Similarly, with both moderate and strong noise, our method generates more true features and less spurious features for varying peak threshold detection values.



Figure 3.9 – Examples of captured imagery, showing a diversity of objects, textures, shapes and sizes. The full dataset contains 20 scenes, each with 22 bursts of 7 images.

We qualitatively demonstrate the feature performance of our method with moderate (top) and strong (bottom) noise levels. Our method outperforms alternative methods of extracting features on a single conventional noisy image and all burst images by detecting all the true features with fewer spurious features in both noise levels compared to alternative methods.

3.4.2 Reconstruction Performance

We demonstrate our method by mounting a monocular machine vision camera FLIR FL3-U3-120S3C-C with an $f/2.1$ lens on a UR5e robotic arm as shown in [Figure 3.1](#). We capture 16 bit raw Bayer images of size 2992×2500 . We capture 22 robotic bursts of 7 images each for a single trajectory as shown in [Figure 3.3](#), and repeat the trajectory for 20 scenes composed of objects with different textures, shapes and sizes in an environment with controlled lighting as shown in [Figure 3.9](#).

In this experiment, we select a trajectory that replicates the motion of a drone mapping the environment at a certain height. However, we demonstrate the performance of this method in more challenging scenes and low-light conditions in [Chapter 5](#) and [Chapter 6](#).

We capture our dataset at 1 ms and 0.1 ms exposure times, adjusting the gain in each case to maximize contrast while avoiding saturation. Apparent motion within each robotic burst is typically about $1/8$ of the camera’s field of view, with examples of faster motion up to $1/4$ of the field of view. We capture a bias frame with lens caps on the camera to block all light. We remove the fixed pattern noise on each image within a robotic burst by subtracting the bias frame from the captured images. This allows us to model the dominant noise on captured imagery as a random Gaussian as discussed in [Chapter 2](#).

We run COLMAP to generate sparse reconstructions of 20 different light-constrained scenes using our methods: burst with merge, burst with merge and Wiener filtering and burst with merge and bilateral filtering. We compare against alternative approaches: burst without merge where we directly use all images in the burst and conventional single-image capture, with the single image corresponding to the center most frame of the burst, as depicted in [Figure 3.2](#).

We build a 4-level coarse to fine hierarchical pyramid with tile size 8 for alignment and use overlapping tiles of the same size for pairwise temporal merge throughout the experiments. Further guidelines on selecting appropriate tile size and pyramid levels for robust merging are discussed in [Section 3.5](#).

We employ COLMAP with default settings, meaning a constant feature peak threshold of 0.0066 is employed for all images. To more fairly compare methods, we also repeat the experiment with peak thresholds adjusted to suit the different noise levels yielded by each method. Because unmerged images have more noise, they call for a more selective peak threshold. We empirically select peak thresholds of 0.001 and 0.005 for the proposed and conventional methods, respectively, as these yield similar levels of spurious feature detection.

Finally, we repeat the same experiments with COLMAP configured to be more permissive of images with few inlier feature matches, allowing as few as 15. This is useful when dealing with challenging scenes and allows more images to be successfully incorporated into the [SfM](#) solution.

Following the feature comparison approach in [20], we evaluate reconstruction performance as shown in Table 3.1 in terms of numbers of keypoints per image, putative feature matches per image, number of putative matches classified as inliers, match ratio: the proportion of detected features yielding putative matches, precision: the proportion of putative matches yielding inlier matches, matching score: the proportion of detected features yielding inlier matches and the mean number of 3D points per image on captured imagery.

In the table, bold highlights the best results, red shows best and second-best results from competing approaches, while green shows best and second-best results from the proposed method. At a moderate noise level, i.e., for images captured over 1 ms, our method shown in green outperforms alternative approaches across all metrics, by reconstructing all images passed for all scenes at default settings. Our method reconstructs hundreds more matches and twice as many 3D points per image compared with alternative approaches.

At strong noise, i.e., for images captured over 0.1 ms, not all images are reconstructed with default settings, and thus, we evaluate by tuning peak threshold values. Our proposed method successfully reconstructs all the scenes with the strongest putative matches per image, inlier matches per image, match ratio, match score and 3D points per image.

3.4.3 Camera Trajectory Accuracy

We evaluate the accuracy of our camera trajectory estimation by employing a robotic arm to collect precise ground truth poses. We align and scale the camera poses estimated through COLMAP with the ground truth poses as there is an inherent scale ambiguity in monocular SfM. We employ Sim(3) Umeyama’s alignment method [119], which effectively accounts for translation, rotation, and scale differences as outlined in EVO implementation ¹, allowing us to calculate a similarity transformation between ground truth and estimated poses. The arbitrary scale factor is established based on

¹<https://pypi.org/project/evo/>

Table 3.1 – Average performance of reconstruction for light-constrained scenes. Bold: Best results, Green: Best/Second best results from our proposed approaches; Red: Best/Second best results from alternative approaches.

Reconstruction Modes	Method	No. of images	No. of images	% images pass	Key Points/Image	Putative Matches/Image	Inlier Matches/Image	Match Ratio	Match Score	Precision	3D Points	3D Points/Image
1 ms Default	Burst without merge	154	152.00	98.70	9.12x10 ³	681	629	7.50x10 ⁻²	6.90x10 ⁻²	0.92	3.98x10 ⁴	258
	Conventional noisy	22	21.60	98.18	9.11x10 ³	656	603	7.20x10 ⁻²	6.60x10 ⁻²	0.92	8.27x10 ³	376
	Burst with merge	22	22.00	100.00	1.05x10 ⁴	866	801	8.20x10 ⁻²	7.60x10 ⁻²	0.93	1.11x10 ⁴	506
	Burst with merge + Wiener	22	22.00	100.00	1.03x10 ⁴	907	842	8.80x10⁻²	8.20x10⁻²	0.93	1.14x10 ⁴	518
0.1ms Default	Burst with merge + bilateral	22	22.00	100.00	1.02x10 ⁴	936	868	9.20x10⁻²	8.50x10⁻²	0.93	1.17x10 ⁴	532
	Burst without merge	154	89.85	58.34	6.19x10 ³	37	29	5.90x10⁻³	4.60x10⁻³	0.78	3.46x10 ³	22
	Conventional noisy	22	13.50	61.36	6.19x10 ³	34	25	5.50x10 ⁻³	4.00x10 ⁻³	0.74	3.82x10 ²	17
	Burst with merge	22	16.55	75.00	9.34x10 ³	45	33	4.80x10 ⁻³	3.50x10 ⁻³	0.73	5.74x10 ²	26
0.1 ms Permissive Matching	Burst with merge + Wiener	22	15.00	68.00	1.10x10 ⁴	47	34	4.20x10 ⁻³	3.00x10 ⁻³	0.72	5.71x10 ²	26
	Burst with merge + bilateral	22	15.80	71.80	7.37x10 ³	50	36	6.70x10⁻³	4.90x10⁻³	0.72	5.66x10 ²	26
	Burst without merge	154	76.90	49.94	6.19x10 ³	37	29	5.90x10⁻³	4.60x10⁻³	0.78	2.62x10 ³	17
	Conventional noisy	22	16.85	76.59	6.19x10 ³	34	25	5.50x10 ⁻³	4.00x10 ⁻³	0.74	4.22x10 ²	19
0.1 ms Peak Threshold Tuned	Burst with merge	22	21.00	95.45	9.34x10 ³	45	33	4.80x10 ⁻³	3.50x10 ⁻³	0.73	7.18x10 ²	34
	Burst with merge + Wiener	22	21.05	95.68	1.10x10 ⁴	47	34	4.20x10 ⁻³	3.00x10 ⁻³	0.72	7.81x10 ²	36
	Burst with merge + bilateral	22	21.90	99.55	7.37x10 ³	50	36	6.70x10⁻³	4.90x10⁻³	0.72	7.80x10 ²	35
	Burst without merge	154	85.95	55.81	5.43x10 ³	43	33	7.90x10 ⁻³	6.10x10 ⁻³	0.77	3.83x10 ³	25
0.1 ms Peak Threshold Tuned	Conventional noisy	22	16.70	75.90	5.43x10 ³	39	30	7.00x10 ⁻³	5.50x10 ⁻³	0.77	5.36x10 ²	24
	Burst with merge	22	20.90	95.00	5.37x10 ³	50	36	9.30x10⁻³	6.70x10⁻³	0.72	8.52x10 ²	39
	Burst with merge + Wiener	22	18.45	83.00	5.98x10 ³	51	37	8.50x10⁻³	6.20x10⁻³	0.73	7.92x10 ²	36
	Burst with merge + bilateral	22	18.70	85.00	7.57x10 ³	53	38	6.90x10 ⁻³	5.00x10 ⁻³	0.72	7.31x10 ²	33
0.1 ms Peak Threshold Tuned + Permissive Mapping	Burst without merge	154	50.70	32.92	5.43x10 ³	43	33	7.90x10 ⁻³	6.10x10 ⁻³	0.77	2.15x10 ³	14
	Conventional noisy	22	19.30	87.73	5.43x10 ³	39	30	7.00x10 ⁻³	5.50x10 ⁻³	0.77	5.70x10 ²	26
	Burst with merge	22	22.00	100.00	5.37x10 ³	50	36	9.30x10⁻³	6.70x10⁻³	0.72	8.64x10 ²	39
	Burst with merge + Wiener	22	21.90	99.55	5.98x10 ³	51	37	8.50x10⁻³	6.20x10⁻³	0.73	8.48x10 ²	39
0.1 ms Peak Threshold Tuned + Permissive Mapping	Burst with merge + bilateral	22	22.00	100.00	7.57x10 ³	53	38	6.90x10 ⁻³	5.00x10 ⁻³	0.72	8.86x10 ²	40

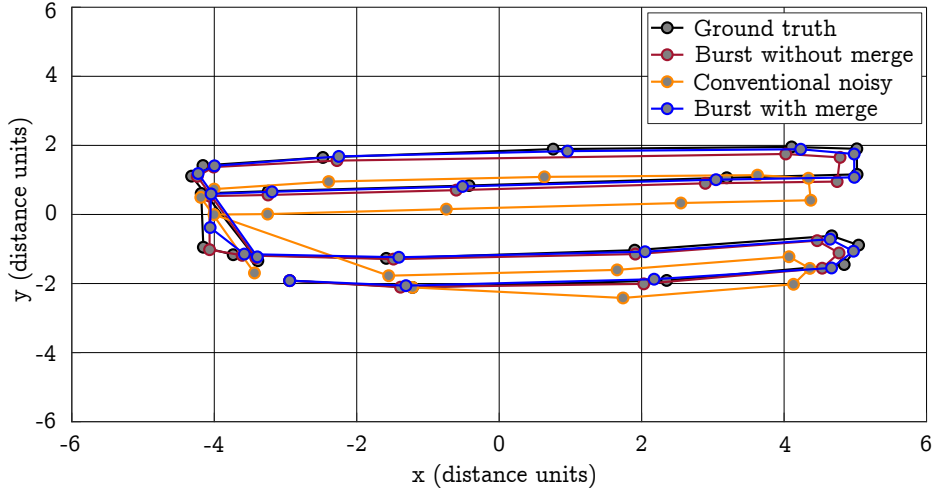


Figure 3.10 – Camera trajectory for a particular captured scene (scene 16) in distance units, a measure between the first pair of registered images in a reconstruction. Our results reconstruct accurate camera trajectories using all input images; the alternative approach that uses reconstructed 94.2% of input noisy images without merge and the conventional noisy approach that uses 95.5% produces less accurate trajectory estimates.

the distance between the initial pair of registered images. Our method reconstructs more accurate camera poses than the competing method demonstrating a trajectory similar to the ground truth as shown in [Figure 3.10](#).

We compute absolute instantaneous error and relative pose error between reconstructed camera poses and ground truth poses for translation and rotation as shown in [Table 3.2](#). The color scheme matches that used in [Table 3.1](#). The green bold values show our method is competitive with moderate noise and outperforms alternative approaches with strong noise by 1cm to 4cm in translation, across all error metrics.

3.4.4 Speed

On an Intel i7-9700 at 4.70 GHz, our MATLAB burst imaging implementation takes 6.54 seconds to align and merge a monochrome burst of 7 images. We expect this could be accelerated substantially. Following align and merge, we employ COLMAP for reconstruction, extracting features using an NVIDIA GTX770. The complete reconstruction pipeline, including time for align and merge, was fastest using our

Table 3.2 – Mean translation error and mean rotational error in camera poses for all reconstruction methods of light-constrained scenes reconstruction methods. Bold: Best results, Green: Best/Second best results from our proposed approaches; Red: Best/Second best results from alternative approaches.

Reconstruction modes		Absolute Instantaneous Error						Relative Pose Error					
		Default (1ms)	Default (0.1ms)	Perm. Mapping	Peak Thresh.	Peak Perm.	Peak Thresh.	Default (1ms)	Default (0.1ms)	Perm. Mapping	Peak Thresh.	Peak Perm.	Peak Thresh.
Burst without Merge	trans. (cm)	0.73	2.83	2.77	2.46	2.75	2.70	2.70	4.28	4.99	3.36	4.66	4.66
	rot. (deg)	0.46	0.72	0.50	0.77	0.49	0.00	0.00	0.05	0.04	0.02	0.04	0.04
	trans. (cm)	0.82	4.51	5.24	5.43	2.92	2.25	2.25	3.90	7.15	5.66	4.71	4.71
	rot. (deg)	0.51	0.78	1.04	1.16	0.62	0.01	0.01	0.03	0.13	0.04	0.09	0.09
Burst with Merge	trans. (cm)	0.74	2.42	2.25	0.98	2.38	1.99	1.99	3.80	3.16	1.99	2.62	2.62
	rot. (deg)	0.45	0.63	0.66	0.34	0.45	0.00	0.00	0.03	0.02	0.01	0.02	0.02
Burst with Merge + Wiener	trans. (cm)	0.79	2.82	2.10	0.89	2.70	2.00	2.00	4.27	3.22	2.29	3.96	3.96
	rot. (deg)	0.45	0.65	0.51	0.42	0.47	0.00	0.00	0.02	0.02	0.01	0.02	0.02
Burst with Merge + Bilateral	trans. (cm)	0.79	2.74	2.81	2.04	1.49	2.64	2.64	3.87	4.24	2.53	2.71	2.71
	rot. (deg).	0.45	0.76	0.65	0.53	0.44	0.00	0.00	0.03	0.03	0.02	0.02	0.02

proposed approach. This is because our method finds fewer spurious features compared to alternative methods, and so fewer putative matches need to be evaluated and rejected. This results in lower overall computational requirements. For a dataset of 154 images in 22 bursts, align, merge, and SfM reconstruction took 3.25 minutes. By contrast, operating on 22 conventional images with no align or merge took 6.32 minutes, and operating on all 154 unmerged images took 49.58 minutes.

3.5 Discussion and Future Directions

We have shown that burst imaging can improve the performance of light-constrained SfM while yielding a net reduction in computational requirements. Although our approach involves increased computation in merging images within a robotic burst, this is offset by the decreased rate of false positive feature detection, reducing the burden of extraneous feature matching and rejection.

The burst approach exhibits multiple trade-offs in acquisition strategy, align and merge parameters, platform motion, and scene content. In the following we offer practical advice for tuning burst imaging for specific robotics applications, based on the experimental findings of this work and on established results from prior works.

- Increase exposure time as high as possible without getting deleterious motion blur [35]. This yields high-SNR input images and ultimately better reconstruction performance.
- Fixed pattern noise can easily become the dominant source of noise in capturing low-light images. This can be addressed by subtracting an average of multiple dark frames taken with the same gain, sensor temperature, and exposure time as the intended images within a burst [120].
- Increase gain as high as possible without yielding excessive saturation [121]. This amplifies both signal and noise, but is important in overcoming the quantisation limit of the camera.

- Maximize the number of images in each burst to maximise SNR in the merged images. The image count is ultimately limited by availability of compute, and the rate of apparent motion of the scene [32]. Motion relative to the common frame should not exceed 1/2 of the total frame. For this reason we also recommend limiting apparent motion relative to the common frame by employing the centre most pose as the common.
- The image pyramids balance computational effort and quality of alignment via the search window size at each pyramid level. Increase pyramid level count and decrease search window size to avoid local minima – this is especially important for images with strong noise [73]. In our experiments, 3 or more pyramid levels maintaining an image size of at least 1/16 of the original image yielded robust alignment.
- For fast motion within the burst, e.g., 1/4 of the total frame or more, increase the search extent to allow for the larger motion. Reduce the image size at each pyramid level – we found it best to go not lower than 1/16 of the original image size. See [33, 35] for more discussion on motion robustness with burst imaging.

In this chapter, we adapted burst photography for robotic vision to enable low-light reconstruction. We enabled the use of direct methods for image registration within the robotic burst and used feature-based SfM to handle the sparsity between robotic bursts for reconstruction. We demonstrated successful reconstruction with decreased failure cases due to non-convergence compared to alternative imaging approaches. We also demonstrated as in Figure 3.1 improved performance relative to conventional imaging in producing high-quality features and feature matches, 3D points per images and accurate camera pose estimation. Our method showed faster reconstruction compared to conventional methods. We expect that in more challenging low-light conditions our method can improve the performance of 3D reconstruction and expand the range in which feature-based reconstruction can be applied.

This work is only a first step toward solving the problem of 3D reconstruction in low light. For future work, we anticipate employing adaptive sampling schemes, in which

the parameters of robotic burst capture and processing are dynamically chosen to suit the situation. We also expect the fusion of complementary sensors to yield interesting results. We discuss this in detail in [Chapter 6](#).

Chapter 4

Burst Feature Finder

*"The true delight is in the finding out
rather than in the knowing"*
- Isaac Asimov

4.1 Overview

We established the viability of adapting burst imaging for light-constrained reconstruction in [Chapter 3](#). However, the effectiveness of this method is ultimately limited by the necessity of aligning tiles between multiple pairs of images in the presence of strong noise. To overcome this limitation, in this chapter, we propose a new approach that exploits the apparent motion within a burst to identify features in a robotic burst. By doing so, we get the maximum [SNR](#) boost possible compared to the previous method which improves feature detection. This is particularly relevant for extremely low-light environments, where the hierarchical tile alignment method used in the burst with merge approach struggles.

In this chapter, we describe an imaging pipeline that captures multiple bursts along the trajectory of a robot and finds features directly within a burst. We describe these features based on their histograms of edge orientations. We validate our method using

real burst imagery collected by a robotic arm. We show our proposed feature extractor finds more higher-quality features and fewer spurious features in light-constrained scenes as shown in Figure 4.1. We evaluate the potential of our method to improve 3D reconstruction in low light and demonstrate our method yielding more complete 3D models, competitive feature performance and more accurate camera pose estimates compared to state-of-the-art feature extraction methods.

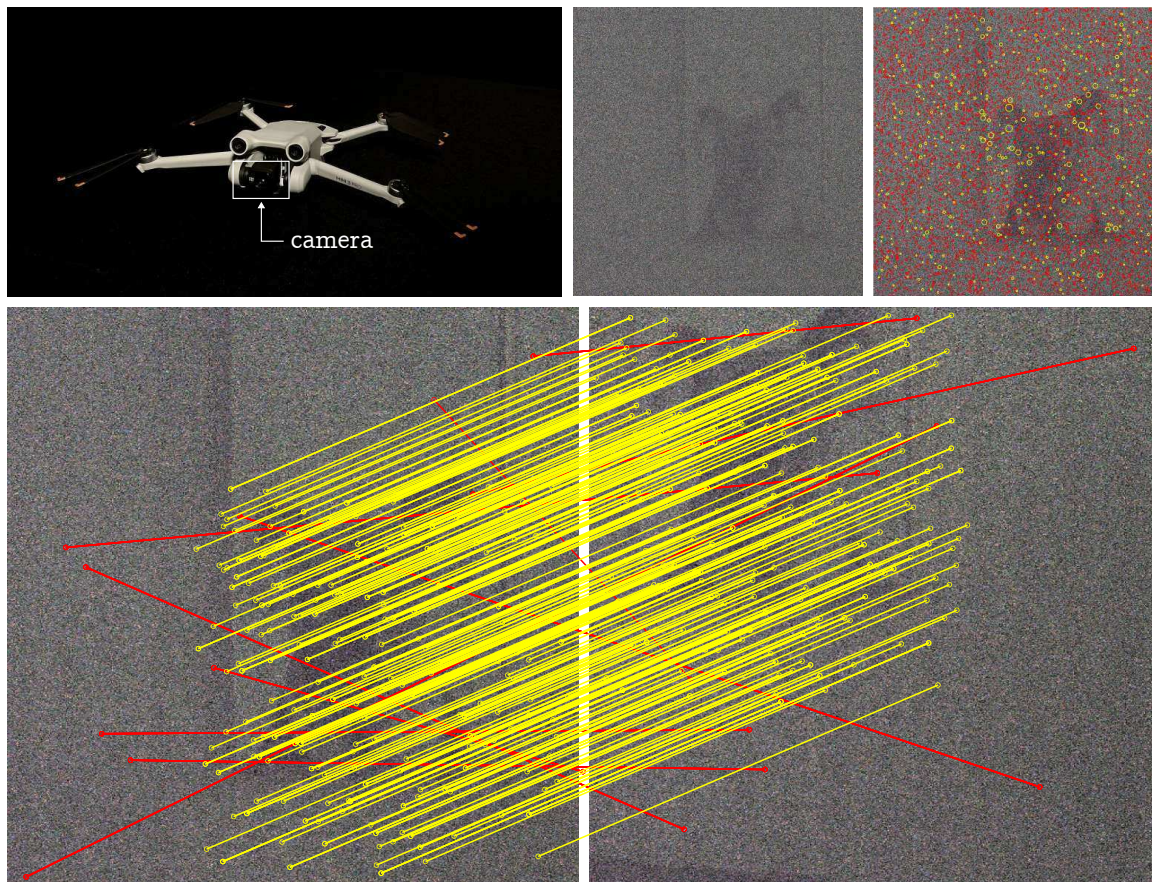


Figure 4.1 – Feature matching in low light: A commercial drone (top-left) captures imagery that is too noisy for conventional 3D reconstruction in low light (top-middle). This is because of the high rate of spurious feature detection and low-quality feature matches offered by conventional features like SIFT (top-right and bottom, red). The proposed BuFF feature yields fewer spurious features, and higher-quality feature descriptors resulting in many more correctly matched pairs (yellow). In this work we show the BuFF feature enables 3D reconstruction in previously prohibitive low-light conditions.

This study aims to identify salient burst features in a robotic burst, with the aim of enhancing light-constrained 3D reconstruction. Parts of this chapter are available in [2] and the code and dataset are available at: <https://roboticimaging.org/Projects/BuFF/>.

In summary, our key technical contributions are:

- We introduce robotic burst feature finder (BuFF), a 2D + time feature detector and descriptor that finds features with well defined scale and apparent motion within a burst of frames,
- We propose the approximation of apparent feature motion as either 1D or 2D linear segments under typical robotic platform dynamics, enabling critical refinements relative to prior work on hand-held imagery, and
- We establish variations of BuFF matched to these apparent motion types and demonstrate that these variants outperform state-of-the-art feature extractors applied on conventional and burst imagery in low-SNR scenes.

4.2 Related Work

Feature-based methods, as discussed in Chapter 2, rely on establishing accurate and reliable pixel-level correspondences across multiple images for visual reconstruction. In robotics, we commonly use SIFT [22] for 2D feature extraction to perform feature-based 3D reconstruction [20].

Feature detection. SIFT builds a difference of Gaussian (DoG) pyramid by convolving a single image with a range of scales of DoG filters. By searching for local extrema in the DoG pyramid, SIFT finds well-defined scale-invariant blob feature locations and describes these features based on their histogram of gradients. SIFT demonstrates higher tolerance to noise and better matching performance comparing to other 2D feature detectors like speeded up robust features (SURF) [40] and oriented FAST and rotated BRIEF (ORB) [41] in reconstruction applications [122, 77].

Drawing inspiration from [SIFT](#) and light field feature ([LiFF](#)) [[117](#)], we find blob features with well-defined scale and apparent motion by directly searching in a robotic burst. Our key insight is that while [SIFT](#) locates blobs with well-defined scales and locations in the 2D image plane and [LiFF](#) deals with identifying blobs with well-defined scales and locations in 3D space on light-field images, our approach leverages the use of multiple 2D images captured rapidly by a monocular camera. By leveraging the 3D motion of the platform, we are able to locate and describe blobs with well-defined scale and apparent motion in the 2D image plane.

Feature description. [SIFT](#) finds correspondences across other images by measuring the similarity between descriptors, and computes feature matches for 3D reconstruction. [SIFT](#) descriptors exhibit robust performance and matching score than other classical feature extractors in the presence of geometric changes for well-lit images [[78](#)]. In our work, we describe each feature similar to [[22](#)] based on the histogram of edge orientations around its local neighborhood. As these descriptors are computed at the stage where images have higher [SNR](#) boost, they have better accuracy and selectivity advantage over conventional methods.

3D Feature extraction. Leveraging [SIFT](#) for video feature extraction, Scovanner et al. [[123](#)] use all the frames in a video for detection and describe the features by extending the [SIFT](#) descriptor to a higher dimension. Another study builds a spatio-temporal space to locate features by explicitly modeling motion using flow methods [[124](#)]. These 3D methods show poor performance [[125](#)] as they rely on individual pixels of low [SNR](#) images in locating local extrema and flow estimates. In contrast, we leverage the inherent motion information of the robotic platform to design a search space with higher [SNR](#) images.

4.3 Robotic Burst Feature Finder

In this section, we extract blob features with well-defined scale and apparent motion on light-constrained robotic bursts and explain its integration into an [SfM](#) pipeline as shown in [Figure 4.2](#).

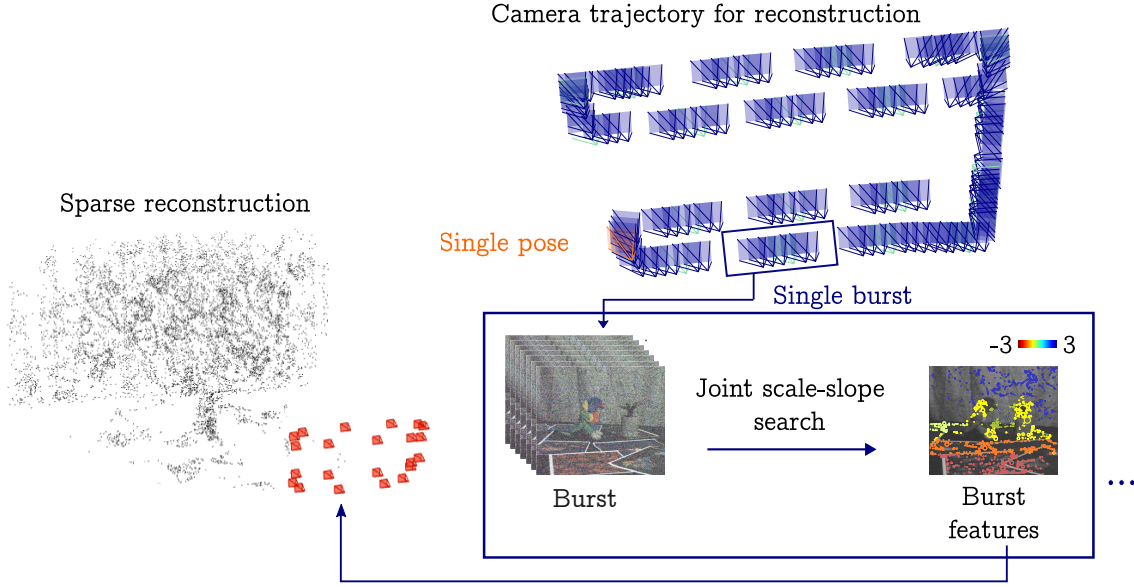


Figure 4.2 – Compared to merging-based approach depicted in Figure 3.3, we directly find blob features with well-defined scale and apparent motion by jointly searching in scale-slope space. We then import burst features into reconstruction pipeline.

4.3.1 Apparent Motion within a Robotic Burst

Capturing burst images using a camera mounted on a moving robotic platform exhibits fundamentally different motion profiles compared to hand-held burst photography. While handshake on handheld cameras contains significant high-frequency components, robotic platform dynamics are usually dominated by higher inertia and thus smoother instantaneous motion. This results in smooth and, for sufficiently fast bursts, locally linear apparent motion.

We identify two main variations in the local motion within a robotic burst as shown in Figure 4.3. Features in a burst captured by a moving robot generally exhibit 2D linear apparent motion, especially when the robot is moving towards or away from the scene. This is because for small motions in a robotic burst, all local motion is well approximated as being linear. In special cases, when a robot is moving orthogonal to the principal axis of the camera, features within a burst exhibit 1D linear apparent motion, i.e., all the features moving in parallel segments in same or opposite directions.



Figure 4.3 – (left): In 3D scenes under general 6-degree of freedom (DOF) platform motion, features in a burst exhibit apparent motion that is well approximated by line segments, even under platform rotation; (right): In the special case of a platform moving orthogonal to the principal axis of the camera, the apparent motion follows parallel line segments. We exploit these observations to search either 2D or 1D spaces of linear apparent motion to detect BuFF features.

We build two variations of the feature finder, BuFF 2D for extracting features in a burst and BuFF 1D as a subset design. Reasons for the design choices are described in Section 4.3.5.

4.3.2 Motion Stack

BuFF identifies 2D + time features in a robotic burst, as shown in Figure 4.4. BuFF 2D searches through location, scale, and a 2D apparent motion space to build the 5D search space and find burst features at unique locations (ϕ), scales (σ), and apparent motions (λ). We pass all the images in a burst, i.e., N frames, through a shift-sum motion filter to generate a motion-filtered image H_M as in,

$$H_M(\phi)|_\lambda = \frac{1}{N} \sum_{n=1}^N I_n(u - \lambda_u \cdot (n - k), v - \lambda_v \cdot (n - k)), \quad (4.1)$$

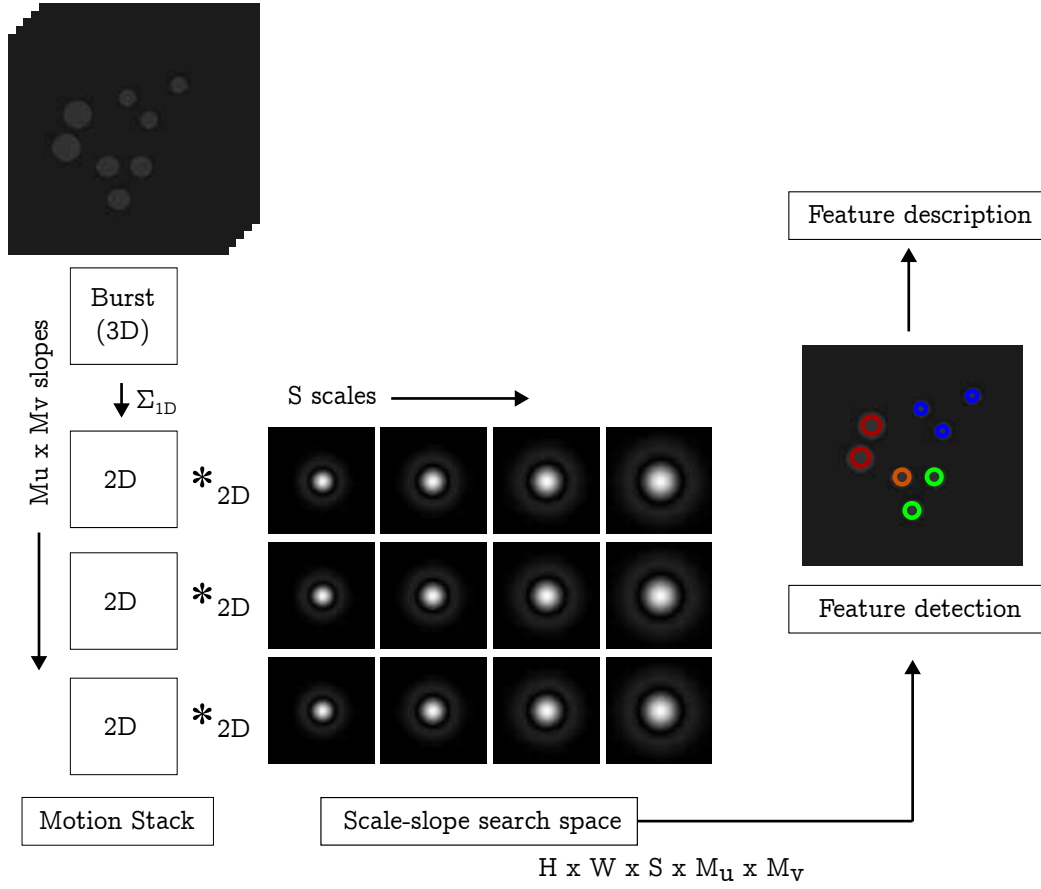


Figure 4.4 – The proposed feature detection architecture: The N -frame input burst is converted to an $M_u \times M_v$ motion-stack by shifting and summing frames across a range of putative apparent motions (“slopes”). This is reduced to an M -frame stack for 1D apparent motion. Each frame of the motion-stack passes through a DoG scale-space filter, and features are detected as extrema in the resulting joint scale-slope search space. Each feature is described using a SIFT-style histogram of gradients applied to its corresponding motion-stack image, rather than the input frames. The resulting features have distinct location, scale and apparent motion, and exhibit high precision, recall, and matching performance.

where I_n is the n^{th} image in a burst and k is the index of the middle frame. $\lambda = [\lambda_u, \lambda_v]$ describes the apparent motion of the pixels in the horizontal and vertical directions and $\phi = [u, v]$ represents the pixel position in the horizontal and vertical directions, respectively.

The motion filter shifts and sums each image in the burst according to the slopes determined by each λ . This process results in a stack of motion-filtered images,

where each image corresponds to a unique slope value. This technique is similar to the design of a focal stack [126].

4.3.3 Scale Slope Search Space

We extend the search space of SIFT by jointly searching for extrema across multiple scales and slopes. To achieve this, we convolve each motion-filtered image, computed by the motion-stack, with DoG filters F_S over multiple scales. This generates a scale-slope search space. We find extrema in this joint 5D search space, D_{5D} as in

$$D_{5D}(\phi, \lambda, \sigma) = H_M(\phi)|_{\lambda} * F_S(\phi)|_{\sigma}. \quad (4.2)$$

Finding features on motion-filtered images rejects spurious features from noisy images. In the special case of BuFF 1D, we consider a linear subspace that corresponds to 1D apparent motion, and this reduces the search space to 4D.

BuFF employs similar parameters as SIFT, such as octaves, levels, contrast threshold, and edge threshold. Additionally, BuFF employs a range of slopes that correspond to apparent motion within a burst. As BuFF finds features within a burst, the number of images in a burst is also a variable.

4.3.4 Descriptor

Similar to SIFT and LiFF, we compute a histogram of edge orientations for each feature which corresponds to a 128-element vector. We describe the features on shift-sum images which have higher SNR compared to the input images in the burst. This allows robust reconstruction of light-constrained scenes. Because we use the images computed from the motion-stack to build the search space, we also have inherent depth information that paves way for other applications including segmentation and depth selection.

4.3.5 Computational Complexity

The computation for BuFF is determined by the number of DoG convolutions, as the cost of computing the motion-stack as in Section 4.3.2 is relatively small and is not considered in our calculation. For BuFF 1D, building a motion-stack prior to convolving with DoG filters is efficient when the number of burst images is higher than the apparent motions. Conversely, for BuFF 2D, convolving burst images with DoG filters and building a motion-stack is more computationally efficient when the number of burst images is less than the square of the apparent motions. By considering the number of burst images as N , slopes as M , and scales as S , the DoG convolution computation for BuFF 1D and 2D can be expressed as either $S \times M$ or $S \times M^2$ with a motion-stack design prior to the DoG stack design, or as $S \times N$ for both with a DoG stack design prior to the motion-stack design.

In BuFF 1D, the motion-stack design followed by the DoG stack design is twice as cost-effective as the reverse order for a 10-image robotic burst with 12 scales and 5 apparent motions. However, in BuFF 2D, using the motion-stack design prior to the DoG stack design is 2.5 times more expensive compared to using the DoG stack design prior to motion-stack design. This is because BuFF 1D convolves 5 motion-filtered images with 12 scales, while BuFF 2D convolves 25 motion-filtered images with 12 scales in the motion-stack design prior to DoG. The optimal BuFF design depends on the range of apparent motion between foreground and background and should be tailored to the specific application.

4.4 Results

In the following, we first evaluate the performance of our feature detector and descriptor on captured light-constrained robotic bursts of a target scene. This is similar to our previous synthetic experiment to validate the feasibility of adapting burst imaging for robotics as discussed in Chapter 3. In this chapter, by capturing an image of the

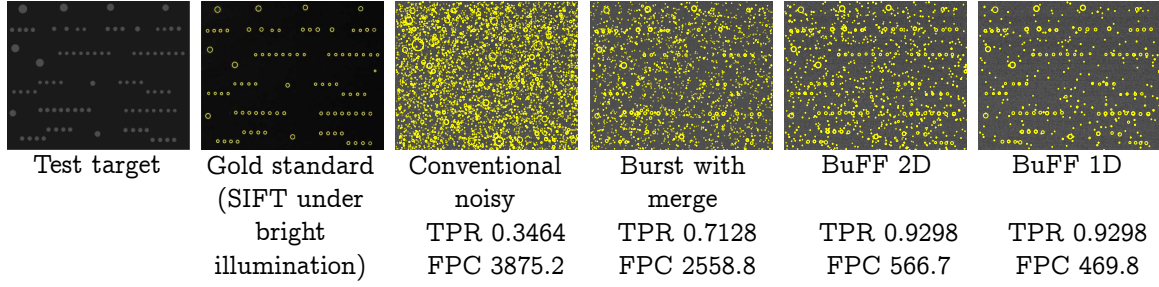


Figure 4.5 – Validation with a printed test target: (left to right) Test target with disks of varying scales to demonstrate feature performance. [SIFT](#) finds features well under bright illumination, but suffers from many false positives and reduced true positives in low light. Employing conventional burst imaging prior to applying [SIFT](#) improves performance, but [BuFF](#) 2D and 1D show much greater performance due to the joint scale-slope search – see also [Figure 4.6](#).

target scene instead of simulating noise, we accurately model the noise of captured imagery for reconstruction.

We use a target scene with varying scales to analyse the nature of feature detection. Then, we quantitatively evaluate the feature performance of [BuFF](#) in an [SfM](#) pipeline across various scenes. We compare against state-of-the-art classical and learning-based feature extractors on conventional images. We also operate state-of-the-art methods on burst-merged images as discussed in [Chapter 3](#). We evaluate feature performance, reconstruction performance and camera pose estimation against alternative methods and demonstrate overall improved performance for light-constrained reconstruction.

4.4.1 Feature Performance in Noise

We print a test target of 90 disks at varying scales with known feature locations to demonstrate feature performance as in [Figure 4.5](#). The color contrast ratio between the disks and the background is 51:26 where 255 is white.

We use a Basler 1600-60um monocular machine vision camera with an f/11 lens to capture bursts of images. The images we capture are 12-bit monochrome and have a resolution of 1600x1200 pixels. To ensure that our images are captured under similar

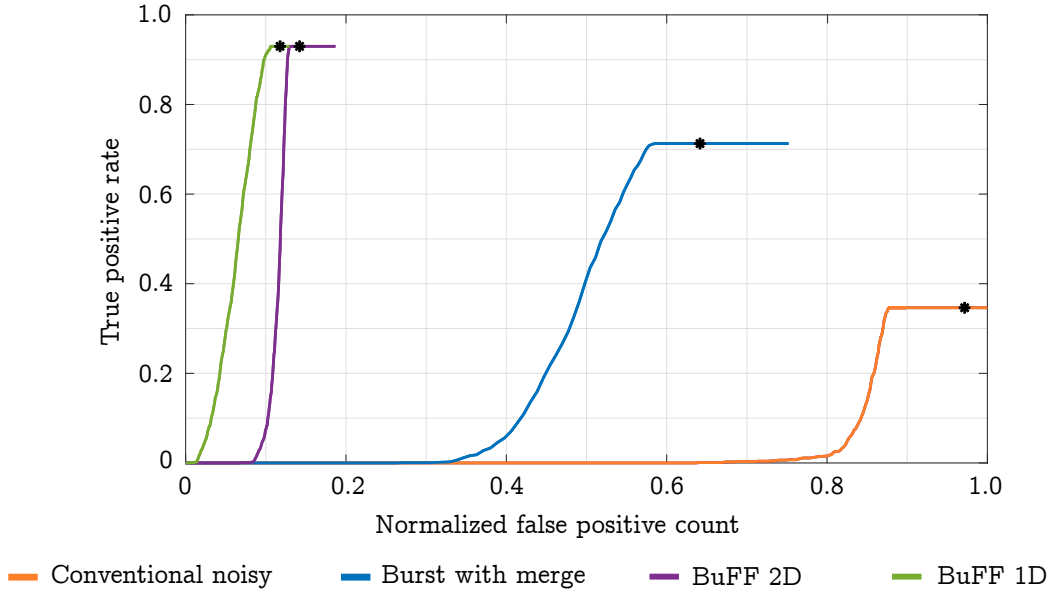


Figure 4.6 – ROC curves for (orange) SIFT on conventional imagery, (blue) SIFT on merged burst imagery, (violet) BuFF with 2D linear local apparent motion, and (green) BuFF with 1D linear local apparent motion. BuFF 1D admits fewer spurious features because of the smaller search space, and both variants of BuFF exhibit higher true positive rates because of the signal boost associated with joint scale-slope search. We use the ROC curves to select comparable peak thresholds for each method, such that 10% of detected features are false positives.

conditions to a clear night sky with a full moon in a well-lit room, we select an appropriate exposure value (EV) to match that of an f/2 camera with an exposure time of 0.16 ms and an illumination of 81.18 lux. This is the same as capturing an image with 5 ms exposure with 1.18 lux of illumination.

We capture the test target scene over a 100 ms exposure time, resulting in a high SNR image, which we use as a gold standard to compare against other images. This gold standard corresponds to capturing the target scene under bright illumination and locates all true features. Our methods both BuFF 1D and BuFF 2D computes the highest overall true features and fewer spurious features on low light robotic burst compared to operating SIFT on both conventional and burst-merged images.

We extract features using our method from 100 captured bursts of the same target scene. We examine the average true positive rate and the average false positive count

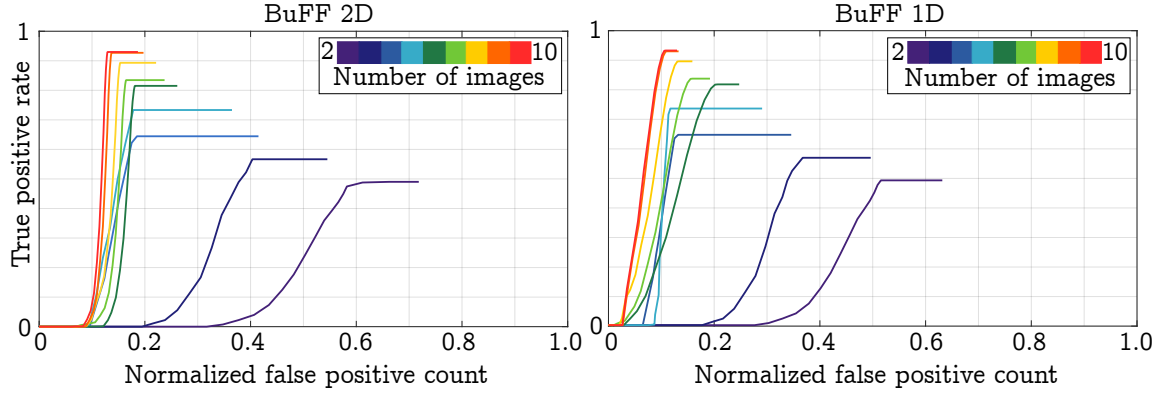


Figure 4.7 – Impact of number of frames in a burst: employing more images improves both spurious feature count and true positive rate. These performance curves vary with camera, scene, and platform motion characteristics. For this scenario, performance saturates at around 10 images per burst.

of the captured imagery of the target scene for a varying peak threshold detection values as shown in Figure 4.6. We compare our methods against the receiver operating characteristic (ROC) of alternative approaches: SIFT on conventional noisy images and SIFT on burst-merged images. We select the peak threshold at which each baseline method performs at its best for the input noise level of the imagery by selecting a peak threshold value at the highest true positive value with 10% of total false positive features.

Conventional methods find fewer true positive features and more spurious features for the captured imagery. Burst-merged images find more true positive features and fewer spurious features compared to conventional methods. Our methods outperform both conventional noisy and burst-merged images finding higher overall true positive features. As the search space is larger for BuFF 2D compared to BuFF 1D, between our methods, BuFF 1D finds fewer spurious features compared to BuFF 2D.

We additionally compare the variation in feature performance of our methods for different numbers of images in a burst as shown in Figure 4.7. As we increase the number of images in a burst, the SNR of the shift-sum images increases by the $\sqrt{N}$ on shift-sum filters. This improves the search in scale-slope space for our feature finder. This gives an advantage for both the detector and the descriptor to find higher-quality features and fewer spurious features during the correspondence search stage

of reconstruction. After nine images, there is minimal difference in extracted higher-quality features and spurious features for the noise level of the captured imagery in a robotic burst. We discuss ways of leveraging this understanding to design adaptive robotic burst imaging in [Chapter 6](#).

4.4.2 Reconstruction Performance

We evaluate reconstruction performance by mounting the same camera under the same lighting conditions as discussed in [Section 4.4.1](#) on a UR5e robotic arm as shown in [Figure 4.8](#). We capture 20 burst sequences with 10 images in each burst for five trajectories. We capture these sequences while the robot is moving towards and away from the scene, resulting in burst sequences having approximately 2D apparent motion between frames. This is similar to both takeoff and landing scenarios of drones during flight.

We also capture bursts having 1D apparent motion between frames in a similar fashion, by capturing multiple burst sequences for five trajectories while the robot is moving orthogonal to the principal axis of the camera. There are a significant number of similar use cases including drone mapping out an area or a mining robot moving along the scene during reconstruction.

Our scene is designed similar to a forest environment containing objects with different textures, shapes and sizes under controlled lighting. We capture our dataset over 5 ms and 100 ms exposure times as noisy and gold standard images respectively to correspond closely to the noise levels of the earlier printed target experiment depicted in [Figure 4.5](#).

We observe and compensate for a special kind of pattern noise – dark current non-uniformity that occurs due to variation from pixels to pixels or columns to columns with dark current generation rates. We remove this by capturing a bias frame under identical temperature and camera parameters as the experiment image with lens caps on the camera to block all light. We remove the fixed pattern noise similar to [Chapter 3](#) on all burst images by subtracting bias frames from the captured images. This

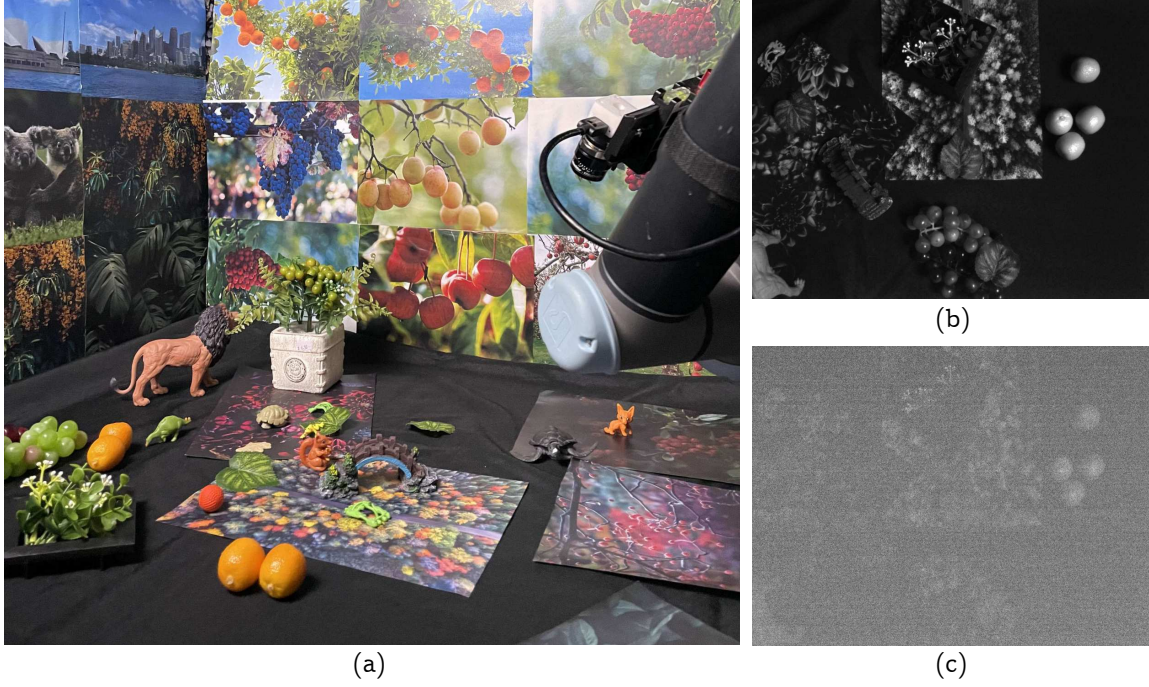


Figure 4.8 – Examples of captured imagery: (a) View of a well-lit scene and arm-mounted camera; (b) a gold standard image captured over 100 ms exposure time; and (c) a single noisy image captured over 5 ms at the same camera pose. The full dataset contains 10 trajectories, each with 20 bursts of 10 images.

allows us to model the dominant noise on captured imagery as a random Gaussian which corresponds to thermal noise.

We compute features from the variants of BuFF and the VLFeat² implementation of SIFT using selected peak threshold as explained in Figure 4.6, edge threshold 10, and DoG scales covering 6 octaves over 4 levels per octave on both conventional and burst-merged images. This is similar to the settings used on the gold standard in the earlier printed test target experiment depicted in Figure 4.5 to generate all the higher-quality features with no spurious features. For the variants of BuFF, we designed the motion-stack over 7 slopes in horizontal and vertical directions accordingly. For the descriptor, we apply L1 root normalization [115] across all methods, similar to [117] to yield improved matches.

²<https://www.vlfeat.org/overview/sift.html>

Following the feature comparison approach in [20], and similar to the evaluation as discussed in Chapter 3, we evaluate reconstruction performance on the captured imagery for both variants of BuFF against state-of-the-art methods for light-constrained reconstruction. Because conventional images outperform alternative imaging approaches of using all frames within a burst in Chapter 3, we evaluate reconstruction using both classical and learning-based methods on conventional images. We also evaluate our method of finding burst features against our previous approach of finding features on a temporally merged image as discussed in Chapter 3. We use SIFT for classical approach as it is the most noise tolerant 2D feature extractor as discussed in Chapter 2 and for learning-based approaches, we extract features and matches using SuperPoint [42] and repeatable and reliable detector and descriptor (R2D2) [44] on our captured light-constrained scenes.

In Table 4.1, we highlight the best results from the variants of our method in green and alternative approaches in red similar to Chapter 3. We compare BuFF 2D for trajectories with 1D motion, and while it uses 92% of all images to reconstruct light-constrained scenes with competitive putative matches, inlier matches and 3D points per image, BuFF 1D subset demonstrate more matches and 3D points overall. This is because of less spurious feature detection in limited search space compared to BuFF 2D.

Similarly, we compare BuFF 1D on 2D trajectories. While it performs competitively with other methods, the accuracy drops because of the constrained motion imposed on 1D motion filters.

Conventional reconstruction using single noisy images failed to identify a good initial image pair and failed to converge during bundle adjustment for both classical and learning-based feature extractors. Our method reconstructs hundreds more matches and 3D points per image compared with burst-merged reconstruction using SIFT.

Our method performs competitively with learning methods. While R2D2 outperforms our method on finding putative matches and therefore match ratio, our method computes more inlier matches and thus more precise matches in light-constrained en-

Table 4.1 – Average performance of reconstruction for light-constrained scenes with 1D and 2D apparent motion. Bold: Overall best results, Green: Best results from our proposed approaches; Red: Best results from alternative approaches.

Apparent Motion	Method	Convergence Rate	Images Pass %	Keypoints/ Image	Putative Matches/ Image	Inliers Matches/ Image	Match Ratio	Match Score	Precision	3D Points/ Image
1D	Gold Standard	1.0	100	2880.96	1.18x10 ⁴	1.17x10 ⁴	4.16	4.14	0.99	257.4
	SIFT	0.0	0	5573.68	6.84x10 ¹	6.41x10 ¹	0.01	0.01	0.87	0.00
	Conventional Noisy	0.0	0	29.48	4.50x10 ⁻¹	0	0.02	0.00	0.00	0.00
	R2D2	0.0	0	2000.00	1.32x10³	1.72x10 ²	0.66	0.09	0.13	0.00
	SIFT	0.8	59	2216.14	3.62x10 ²	3.32x10 ²	0.19	0.17	0.93	31.6
	SuperPoint	1.0	69	154.64	1.26x10 ¹	7.45x10 ⁰	0.58	0.29	0.56	4.20
	R2D2	0.6	37	2000.0	1.20x10 ³	2.18x10 ²	0.60	0.11	0.18	3.20
	BuFF 1D	1.0	98	1386.30	3.84x10 ²	3.75x10²	0.28	0.27	0.98	36.50
	BuFF 2D	1.0	92	1635.82	3.45x10 ²	3.38x10 ²	0.23	0.23	0.98	36.20
	Gold Standard	1.0	100	3270.16	1.30x10 ⁴	1.29x10 ⁴	3.89	3.87	0.99	305.50
	SIFT	0.0	0	5680.10	1.84x10 ¹	1.43x10 ¹	0.00	0.00	0.86	0.00
	Conventional Noisy	0.0	0	25.88	3.38x10 ⁻¹	0	0.01	0.00	0.00	0.00
	R2D2	0.0	0	2000.00	1.30x10³	1.88x10 ²	0.65	0.10	0.15	0.00
2D	SIFT	0.8	77	2745.74	4.85x10 ²	4.57x10 ²	0.18	0.17	0.95	32.20
	SuperPoint	0.4	19	94.40	2.87x10 ¹	8.15x10 ⁰	0.24	0.07	0.24	10.84
	R2D2	0.4	41	2000.00	1.25x10 ³	1.95x10 ²	0.62	0.10	0.16	3.99
	BuFF 1D	0.8	85	2572.40	5.55x10 ²	5.43x10 ²	0.21	0.21	0.95	42.20
	BuFF 2D	1.0	98	2143.96	5.65x10 ²	5.51x10²	0.26	0.26	0.98	47.20

vironments. Our methods also compute more 3D points with higher convergence rate compared to other classical and learning methods.

4.4.3 Camera Trajectory Accuracy

We compute average absolute trajectory error and average relative pose error for translation and rotation from the poses estimated by COLMAP as shown in Table 4.2. The color scheme matches that used in Table 4.1.

The results show the two BuFF variants outperforming burst with merge in translation and BuFF 2D showing a slight advantage over BuFF 1D for rotation estimates. The variants of BuFF outperforms all the other methods, including R2D2, overall for both cases with apparent motion of 1D and 2D. Because conventional noisy methods fail to converge throughout, there are no camera pose estimates for that method.

We demonstrate the camera pose estimates qualitatively as in Figure 4.9. We align and scale the camera poses estimated by COLMAP to the ground truth trajectory poses collected by the robotic arm by operating Sim(3) Umeyama’s alignment method [119] similar to Chapter 3 as there is an arbitrary scale factor involved in monocular SfM. While relaxing inlier rate settings encourages conventional methods to converge, the camera pose estimates are inaccurate. Both SuperPoint and R2D2 fail

Table 4.2 – Mean translation error and mean rotational error in camera poses for reconstruction methods of light-constrained scenes with 1D and 2D apparent motion. Bold: Overall best results, Green: Best results from our proposed approaches; Red: Best results from alternative approaches.

Apparent motion	Method	Absolute trajectory error		Relative pose error	
		trans.(cm)	rot.(deg)	trans.(cm)	rot.(deg)
1D	Burst with Merge - SIFT	2.84	2.07	4.97	0.73
	Burst with Merge - SuperPoint	2.13	2.17	5.12	0.81
	Burst with Merge - R2D2	1.71	1.20	3.17	0.04
	BuFF 1D	1.23	1.24	2.00	0.05
	BuFF 2D	1.27	0.62	2.01	0.02
2D	Burst with Merge - SIFT	1.80	1.11	3.60	0.04
	Burst with Merge - SuperPoint	1.56	1.06	3.71	0.04
	Burst with Merge - R2D2	1.48	0.55	2.84	0.04
	BuFF 1D	1.53	1.17	2.99	0.05
	BuFF 2D	1.44	0.58	2.55	0.03

to converge for this particular scene with relaxed settings. Estimating camera poses using burst-merged images across all methods show poor estimates in translation. Both variants of our method compute accurate camera pose estimates compared to the competing methods using both classical and learning-based feature extractors on single images and burst-merged images.

4.4.4 Speed

For a dataset of 200 images in 20 bursts captured over 5 ms with strong noise, our MATLAB implementation for detecting and describing burst features and reconstruction BuFF 1D took 18.42 minutes and BuFF 2D took 108.5 minutes on an Intel i7-9700 at 4.70 GHz as in Table 4.3. The speed of the implementation depends on the noise levels of the images, as noisy images produce spurious features and describing each detected feature requires a considerable amount of time.

The reported time for BuFF 2D and BuFF 1D followed the design of building a motion-stack prior to the DoG convolution stage, which is the slower of the two options for BuFF 2D and faster for BuFF 1D as discussed in Section 4.3.5. In comparison, operating the SIFT MATLAB implementation on 20 conventional noisy images took 15.04 minutes, while operating on burst-merged images with align and merge took 13.86 minutes. Learning methods that used pre-trained models and corresponding evaluation showed significantly faster performance. Note, all the learning methods are accelerated with NVIDIA GeForce RTX 3080 Laptop GPU. We expect

Table 4.3 – Time taken for reconstruction including feature extraction.

Method		Time taken (min)
Conventional Noisy	SIFT	15.04
	SuperPoint	0.32
	R2D2	2.00
Burst with Merge	SIFT	13.86
	SuperPoint	0.27
	R2D2	1.56
BuFF 2D (Ours)		108.50
BuFF 1D (Ours)		18.42

BuFF implementation to be substantially accelerated with an adaptive design and an

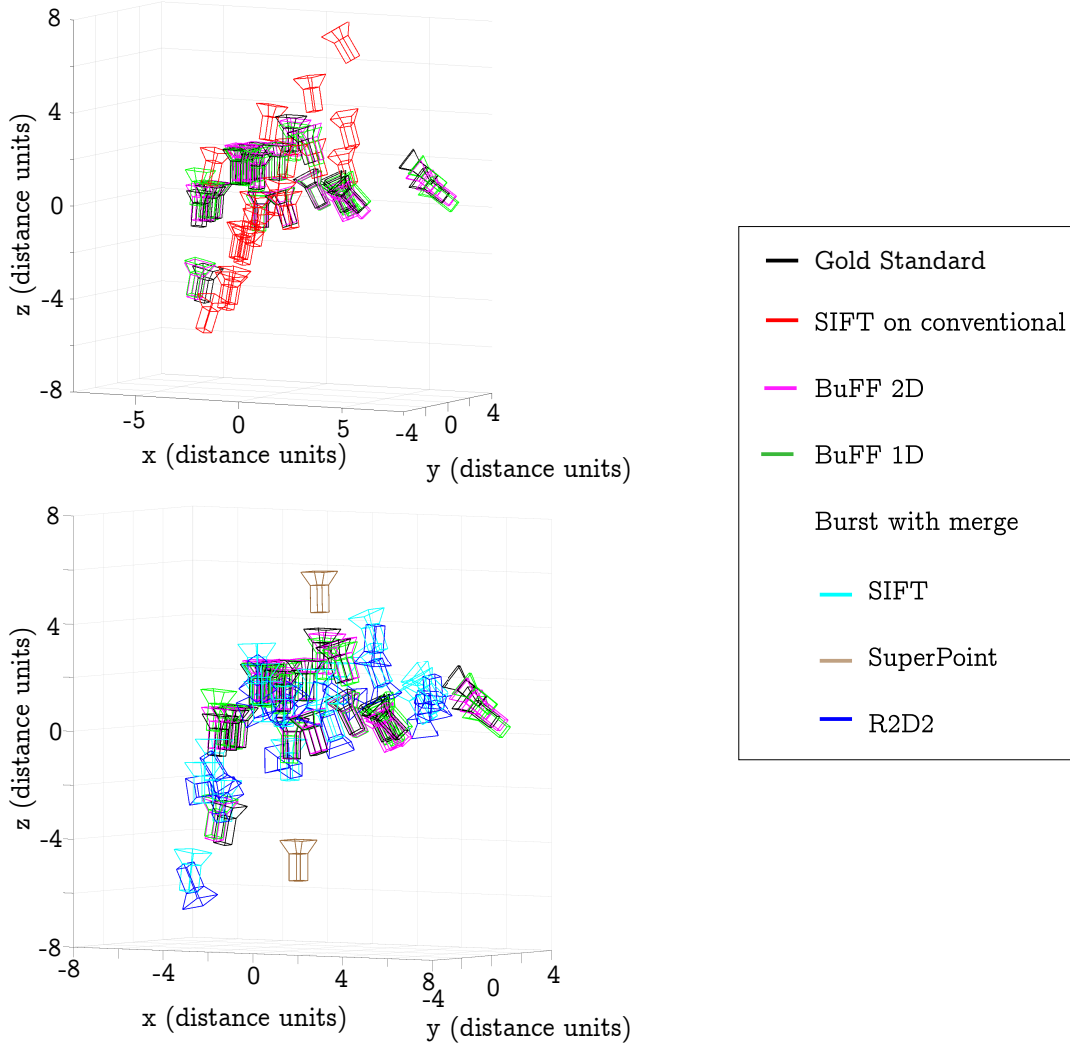


Figure 4.9 – Localisation performance: (top) Conventional imagery using SIFT yields very poor camera pose estimates (red) and even fails to converge without relaxing inlier rate settings. (bottom) Employing burst-merged images using SIFT (cyan) reconstructs 90% of input images but with poor pose estimates (orange). Extracting features and matches using SuperPoint (brown) and R2D2 (blue) reconstructs 10% and 85% of input images respectively but continues to estimate inaccurate poses. The proposed BuFF variants successfully employs the most images and yields the most accurate pose estimates, with the 2D variant (violet) showing slightly lower accuracy than BuFF 1D (green), due to increased spurious features associated with the larger 2D search space.

optimized compiled language, with potential for further computation improvement through parallel architecture of GPUs.

4.4.5 Failure Cases

The performance of the approach is influenced by the level of signal present in the captured burst. Having more images in the burst (N-frames burst) benefits feature detection, but as $\sqrt{N}$, where N is the number of images in a burst. Ultimately, there needs to be enough signal in the images for BuFF to detect features, which is particularly challenging in low-contrast scenes.

Our methods also struggle to perform on bursts of images that display a wide range of apparent motion. This causes aliasing features which could potentially be eliminated by employing faster bursts and enabling smoother parameterization over apparent motion count. Examples of failed cases with larger apparent motion between frames and extremely low contrast are shown in Figure 4.10.

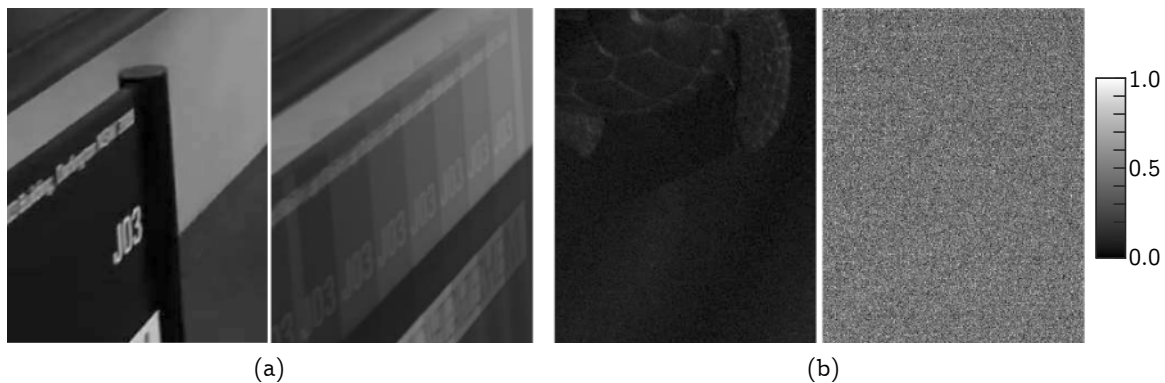


Figure 4.10 – Failure cases: (a, left) A single input frame and (right) an image from the motion-stack with inaccurate apparent motion values show aliasing effect. Employing faster bursts and reducing the slope intervals eliminate this kind of aliasing. (b, left) A low-contrast textured scene in bright illumination, and (right) the same under low light showing insufficient signal for feature detection. Ultimately, sufficient signal is necessary for BuFF to identify features. (Contrast-adjusted image for visualisation)

4.5 Discussion and Future Directions

In this chapter, we introduced a novel robotic burst feature finder to reconstruct low-light scenes. We captured multiple-bursts sequences along typical motion trajectories and detected burst features in a scale-slope search space. We employed a descriptor that provided additional advantage with higher SNR boost compared to conventional methods.

We demonstrated improved performance compared to conventional imaging and burst-merged reconstruction using SIFT in terms of higher-quality features, fewer spurious features, better convergence rate, improved match score, higher match ratio and precise 3D points per image. We also compared against state-of-the-art learning-based feature extractors and demonstrated competitive performance across feature matches. We showed more complete 3D models with the highest 3D points and convergence rate. We also showed overall improved camera pose estimates on conventional images and burst-merged images for low-light scenes. Between our two burst methods for reconstruction of light-constrained scenes, we demonstrated the benefits of directly finding burst features compared to 2D feature extraction on temporally merged robotic bursts.

We expect that this work can improve 3D reconstruction in low light, allowing robots to perform better in challenging mining missions and drone delivery at night. For future work, we anticipate learning-based approaches for burst feature extraction and object tracking and reconstruction of scenes obscured by visually challenging features like snow. We also expect complementary sensors to aid in improving reconstruction in low light – see Section 6.2 for more details.

Chapter 5

Learned Burst Feature Finder

*"The imagination of nature is far, far
greater than the imagination of man"*

- Richard P. Feynman

5.1 Overview

We presented two methods of allowing robotic burst imaging schemes as input for the correspondence search stage in light-constrained reconstruction. We demonstrated the viability of adapting burst imaging for reconstruction in [Chapter 3](#) and demonstrated the ability of burst with merge compared to alternative imaging approaches. We deployed a robotic burst feature finder, outlined in [Chapter 4](#), to jointly search for burst features in scale-slope space. However, physics-based approaches face limitations in locating salient features and describing them accurately in extremely low-light scenes with millilux illumination. This is because of the limited information available in any noisy signals for feature localization and the further degradation caused by quantization.

In this chapter, we present a learning-based pipeline for robotic burst imaging that utilizes joint detection and description to locate sparse features within a burst. By

leveraging the use of metric learning loss from existing 2D feature extractors [44, 127] and adapting for robotic burst, we identify robust features for reconstruction in light-constrained bursts, effectively avoiding regions with low contrast, featureless texture, and indistinct features that could hinder appearance-based detection and description. By directly optimizing average-precision (AP) for descriptor computation, we attain high-quality true features that directly enhance feature matches.

To evaluate, we initially compare our proposed approach with the single-image-based R2D2 [44]. We employ HPatches dataset [128] with varying illuminations and viewpoints to generate synthetic robotic bursts. We demonstrate improved performance for feature matching and camera pose estimates using the widely available COLMAP SfM software [20] with our learned burst features. This highlights the performance improvement in reconstruction achieved by extending learning-based 2D feature extraction to burst imagery under low-light conditions.

We validate our method using BuFF dataset – captured imagery as discussed in Chapter 4 and captured burst imagery collected by DJI Mini Pro 3 and DJI Phantom Pro 4 drones operating in extremely low-light conditions. We compare our proposed approach against state-of-the-art physics-based and learning-based techniques on conventional images. We also compare against other robotic burst imaging schemes as described in Chapter 3 and Chapter 4. We demonstrate higher true features with fewer spurious features, leading to better quality matches for scenes captured in millilux illumination as shown in Figure 5.1. We show overall improved reconstruction performance, resulting in more complete 3D models.

This work aims to identify learning-based features in low-light robotic bursts that enhance low-light reconstruction, benefiting a wide range of drone applications, including nighttime drone delivery and understanding nocturnal behaviors of animals.

In summary, our key technical contributions are:

- We introduce a learning-based robotic burst feature finder, a joint feature detector and descriptor that finds learned features with well-defined scale within a light-constrained robotic burst,

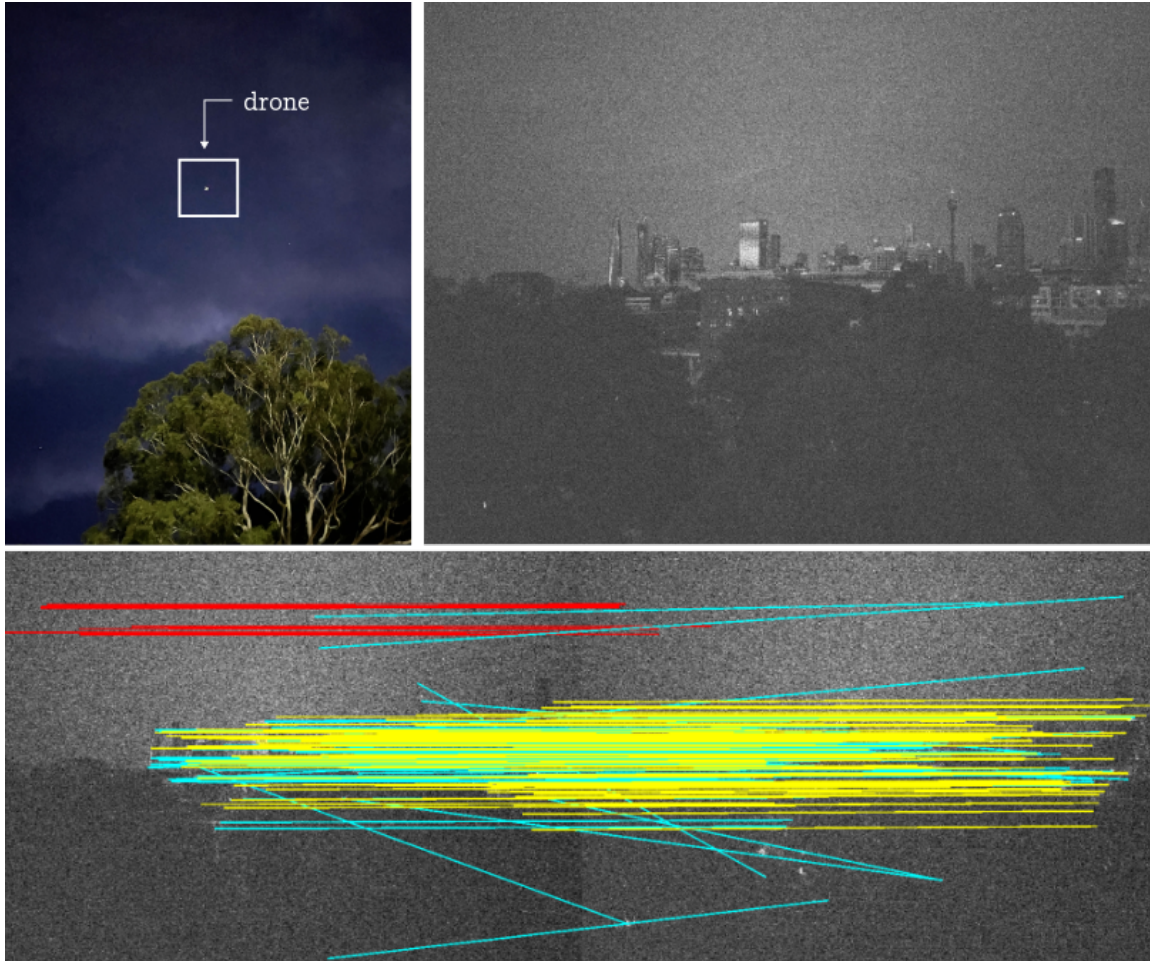


Figure 5.1 – Feature matching at night: A commercial drone (top-left) captures imagery at millilux conditions at night that is too noisy for conventional 3D reconstruction in low light (top-right). This is because of the high rate of spurious features offered by conventional features like **SIFT** (blue) and less reliable matches offered by learned features like **R2D2** (red) on noisy images. The proposed LBurst yields higher-quality feature matches from reliable regions of the images (yellow) for reconstruction. In this work we show that our feature finder enables 3D reconstruction in millilux conditions for existing commercial drones.

- We enable robots to detect noise-tolerant features in low light by employing consistent motion within bursts and leveraging uncorrelated noise between them, and
- We demonstrate overall improved 3D reconstruction from drone images in low-light (millilux) conditions, outperforming conventional and learning-based feature extraction and robotic burst feature finder in low-SNR scenes.

5.2 Related Work

In this thesis, we discussed feature extraction methods for reconstruction using conventional images in [Chapter 2](#), highlighting the role of physics-based approaches in state-of-the-art feature extraction techniques in [Chapter 3](#). We extended it to 3D feature extraction methods and demonstrated a physics-based technique that finds features in a robotic burst. In this section, we discuss the role of machine learning in producing pixel-wise correspondences for 3D reconstruction.

The early stages of learning aim at detecting features, imitating hand-crafted features [\[129, 130, 131\]](#), or involve training a classifier to predict keypoints that may not be useful when matching them during reconstruction [\[132\]](#). These methods remain limited by the poor accuracy of keypoint detection on light-constrained images.

The use of convolution neural network (CNN) [\[133, 134\]](#) allows direct mapping of raw pixels to locations of features through a series of non-linear processing layers, eliminating the need for hand-crafted features. While learning-based detectors demonstrated better performance compared to classical methods such as SIFT, especially in recovering relative orientation [\[135, 136\]](#), hand-crafted descriptors outperformed learned descriptors in image-based reconstruction [\[137\]](#) up until recent years.

Latest studies formulate matching as a problem of metric learning [\[138, 127, 44\]](#), in which the model learns a distance metric by optimizing a loss function that encourages similarity between image patches. The loss function varies as pairwise loss function to compare between two patches [\[139, 140\]](#), triplet losses for three patches [\[141, 142\]](#)

and global losses [143, 144] for more than three patches and this patch similarity reflect Euclidean distances. This substantially improves matching performance for reconstruction [78] compared to classical methods.

Similar to L2Net [144] and HardNet [142], descriptors optimized for average-precision (DOAP) [127] adopts a similar framework but aims to train the model to maximize the mean average-precision (mAP) by directly optimizing the area under the precision-recall curve. This directly improves feature matching performance. In order to make the AP calculation differentiable, the study replaces the traditional discrete histogram with a linear interpolation, resulting in state-of-the-art performance for reconstruction using learning methods.

Taking an end-to-end approach, learned invariant feature transform (LIFT) [145] utilizes three separate CNN components that feed into each other. SuperPoint [42] tackles end-to-end as a supervised task with self-labeling, where an initial keypoint detector is pre-trained on a large dataset of millions of synthetic geometric shapes and automatically generates pseudo-ground truth keypoints. The dependency on pseudo-ground truth keypoints could limit the reliability of the extracted features, presenting a potential bottleneck within the process. D2-Net [43] locates 2D keypoints through joint detection and description using a single CNN with shared weights, while R2D2 outperforms D2-Net and other studies [42, 145, 146] in matching performance by considering reliability as an attribute during detection.

We visualize spurious feature detection associated with strong noise when employing R2D2 on a toy image with low SNR, as shown in Figure 5.2. In the color scheme employed for the confidence maps, red denotes the lowest confidence levels while green shows the highest confidence levels. The color scheme remains consistent across all figures within this chapter. The low confidence associated with both repeatability and reliability confidence maps highlight the need for existing state-of-the-art techniques to deal with strong noise which is common among images captured in extremely low-light conditions.

Inspired by the success of multi-image strategies [37, 36] in addressing visually challenging images, we adapt R2D2, a single-image-based feature extractor, for robotic

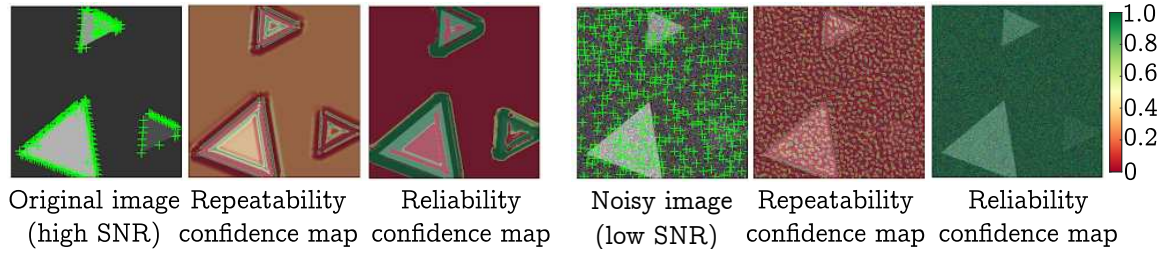


Figure 5.2 – We apply the state-of-the-art [R2D2](#) to a toy image with a high [SNR](#) (left) and demonstrate the corresponding keypoints and reliable regions within the image. However, when dealing with a toy image of low [SNR](#) (right), feature localization results in spurious feature detection, accompanied by inaccurate confidence maps that affect repeatability and reliability scores. The confidence maps employ a color scheme where red denotes the lowest confidence, while green represents the highest confidence. Keypoints are determined based on the high-scored product of repeatability and reliability scores associated with each pixel location.

burst imaging. We tailor the input structure to accommodate a burst of images, enabling the network to utilize temporal information during training. By exposing the network to the typical noise profile and motion variations commonly encountered within a robotic burst, our approach aims to detect sparse features that remain robust regardless of the noise or motion characteristics present within the robotic bursts.

5.3 Learned Robotic Burst Feature Finder

We locate learned features with well-defined scale within low-light robotic bursts using a self-supervised approach enforcing burst pair similarity as described in [Section 5.3.1](#). By employing a flow map with known transformation between common images (i.e., the middle frames of the bursts) of the ground truth bursts during training, the network learns to find noise-tolerant features. These features are compact 128-dimensional descriptors, obtained using [mAP](#)-based loss similar to [R2D2](#) [44] and [DOAP](#) [127]. We select keypoints based on detection and description scores, applied across scales to identify multi-scale features. During inference, we compute multi-scale feature locations and descriptors within a single burst, representing the sparse pose of the common image. For reconstruction, we extend this over multiple bursts captured along a trajectory as shown in [Figure 5.3](#). The following outlines the

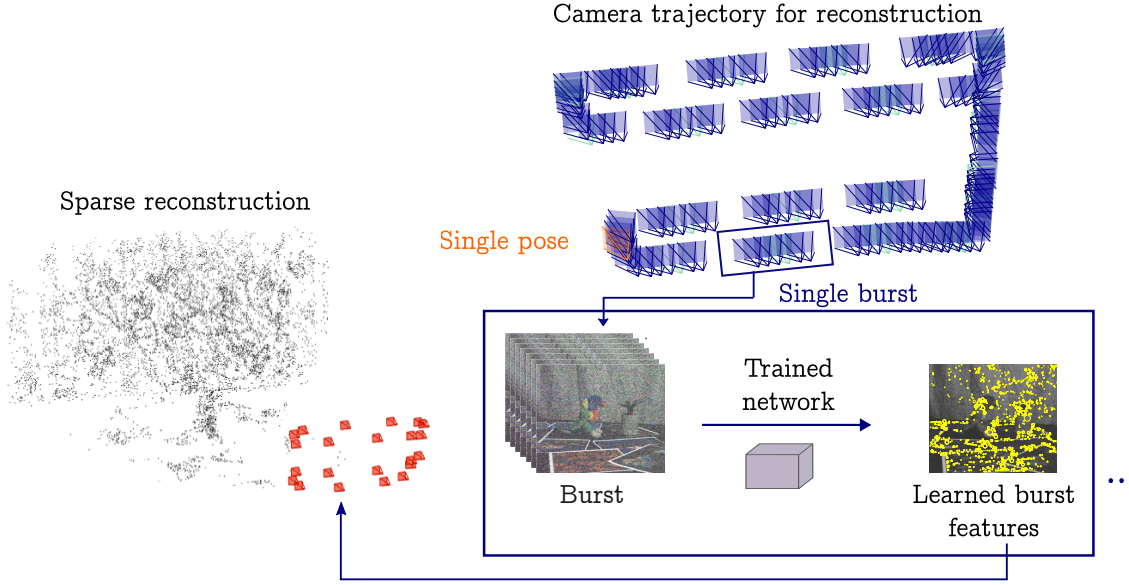


Figure 5.3 – Comparing to the merging-based approach depicted in Figure 3.3, and locating physics-based burst features as described in Figure 4.2, We pass each captured burst through our trained network and locate features in multiple scales. We compute corresponding matches between bursts and perform 3D reconstruction.

joint architecture for detection and description and the burst generation pipeline for training and inference.

5.3.1 Joint Learning of Detection and Description

We employ a fully-convolutional network with L2-Net [144] as the backbone architecture, similar to R2D2 [44] to directly map raw pixel data from a burst of images to multiple outputs, including the sparse keypoint location, pixel-wise descriptors, and their corresponding confidence maps as in Figure 5.4. Because feature-based SfM can handle large occlusions between the bursts, we focus on finding features of the common image that represents the burst, i.e., the middle frame in the case of a robotic burst [1]. The burst has dimensions of $H \times W \times N$, where $H \times W$ represents the size of each image, and N denotes the number of images in the burst.

We train our model by leveraging a pair of robotic bursts and a flow map which leverages the known transformation between the common images of the bursts as

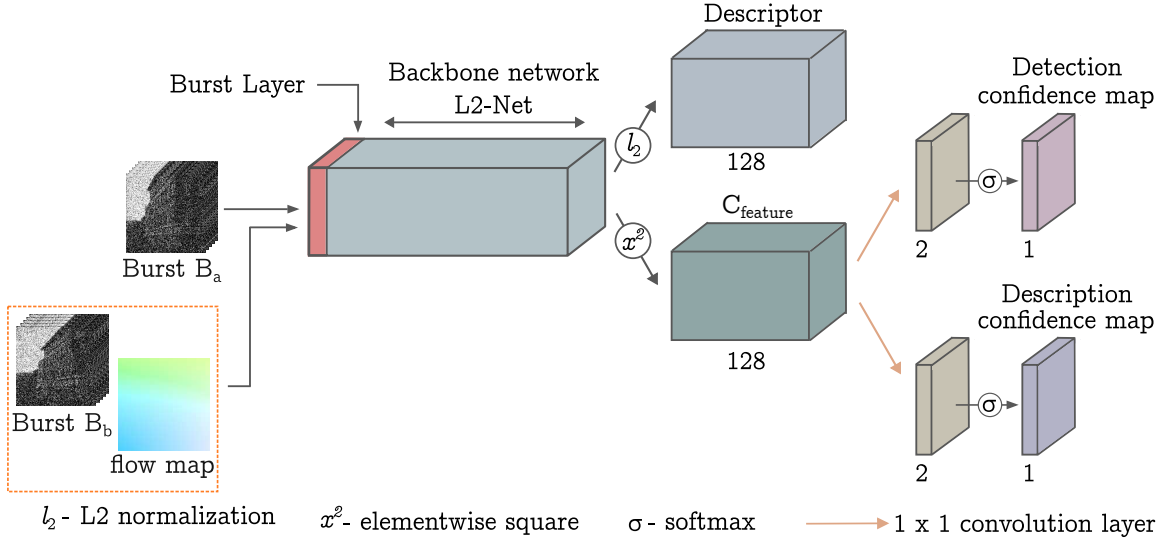


Figure 5.4 – LBurst architecture employs a pair of robotic bursts and a flow map with the known ground-truth transformations between the common images (i.e., the middle frames) during training. A specialized burst layer handles burst image processing. The architecture employs a fully convolutional L2-Net, similar to R2D2, generating a 128D descriptor with L2 normalization and a squared element-wise operation yielding $C_{feature}$. Following a 1×1 convolution for dimension reduction, a softmax operation computes detection and descriptor maps.

described in [Section 5.3.2](#). We pass the burst of images through an adaptable burst layer designed to handle multiple input images. The intermediate layers allow for non-linearities and normalized activations to maintain the input spatial resolution by enabling dilated convolutions and replacing the last 8×8 convolution layer with three successive 2×2 convolution layers, similar to the single-image based approach in [R2D2](#). In contrast to [R2D2](#), by incorporating the flow map that employ the known local transformation between the common images, we impose a spatial pose constraint while integrating temporal information into our architecture. This allows us to capture and leverage the temporal relationships in burst images.

We compute per-pixel 128-dimensional descriptors with $L2$ -normalization on the resulting tensor from the backbone. This normalization is beneficial for [CNN](#) descriptors for image matching tasks [[139](#), [144](#)]. We also apply an element-wise square operation to generate an intermediate $C_{feature}$ representation from the output tensor of modified L2-Net. Passing through an additional 1×1 convolution layer to reduce

the number of channels and a softmax operation to generate a probability distribution over the pixels, we generate confidence maps corresponding to detection and a description.

We generate the detection confidence map by maximizing the cosine similarity between pairs of patches from the common images, and thereby identifying the local maxima on the common image. Because the noise in low SNR images are random in both spatial and temporal domain, locating features based on similarity avoids spurious features. We average the cosine similarity loss L_c over each small patch over the common image and additionally incorporate a local peakiness loss L_p inspired by [44] to enhance the detection of features in low contrast regions of low SNR images. This loss computation promotes noise tolerant features within a burst and consistency of features between the bursts. The overall detection loss L in a robotic burst B_a is computed as,

$$L(B_a, B_b, U_i) = L_c(B_a, B_b, U_i) + \frac{L_p(B_a) + L_p(B_b)}{2}, \quad (5.1)$$

where B_b represents bursts of images transformed by the flow map U_i , ensuring alignment between the common images of B_b and B_a .

We aim to maximize the AP to evaluate the reliability of the patches in the description confidence maps. By doing so, we improve the likelihood of identifying true features while limiting the presence of spurious features, particularly in low SNR images. By excluding non-distinctive regions during reconstruction, we further reduce the presence of unreliable features that have lower matching probabilities in the reconstruction pipeline. The confidence maps, ranging from 0 to 1, indicate the presence of local maxima for a specific patch size in the detection confidence map and represent the highest likelihood for matching in the description map.

5.3.2 Robotic Burst Generation for Training

We generate a pair of robotic bursts using a single image such that each burst shares the same scene with homographic transformation as if captured from two different

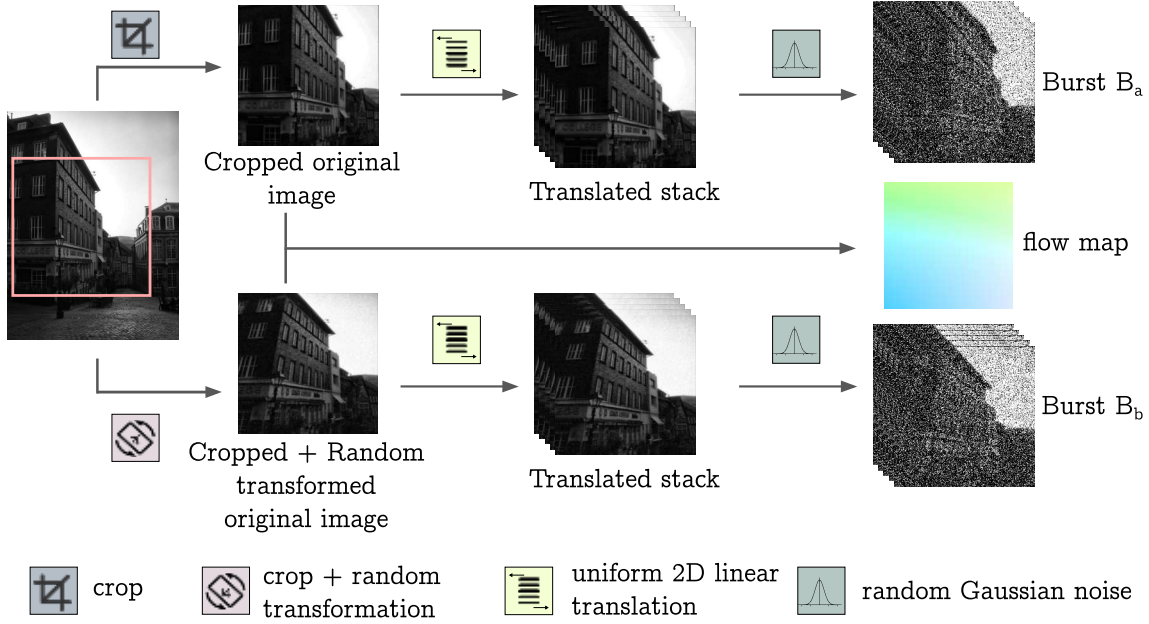


Figure 5.5 – Overview of the synthetic robotic burst generation for model training. Cropping and transforming a single image to yield a pair of perspective-altered scenes, and by applying uniform 2D translations within each burst we generate a pair of burst images. We introduce random Gaussian noise to each image in a burst and employ flow maps with known transformation to ensure accurate vector field between common images (i.e., the middle frame within the generated bursts).

angles as in Figure 5.5. This allows the network to understand the variations in motion temporal information between bursts. We implement consistent uniform 2D translations within individual bursts, while introducing random translations between the bursts. For every image in a robotic burst, we introduce random Gaussian noise to simulate thermal noise present in low-light conditions. To establish the relationship between the common image of each burst, we generate a flow map U_i based on the known transformations. This flow map represents the ground truth spatial correspondence between the common images of burst B_a and B_b and thus learns to disregard the motion within robotic bursts – see further discussion in Section 5.4.1.

Our architecture learns to handle various challenges encountered on robotic bursts during inference, such as local occlusions, parallax motion, and warp artifacts even though the model is not explicitly trained on these specific attributes. This is because we use local cosine similarity over small patches for feature detection, and average

them over all the patches and by doing so, we treat apparent motion associated with the patches locally.

5.3.3 Robotic Burst Generation for Inference

When a drone captures multiple bursts of images using a monocular camera, it usually exhibits 2D linear motion among burst frames as discussed in [Chapter 4](#). In addition, the noise levels in the captured images captured in millilux illumination, typically follow a random Gaussian distribution [112]. We design our burst generation process for training to closely emulate the motion and noise characteristics of captured drone burst imagery for reconstruction. We also generate low-light robotic burst for synthetic experiments, closely imitating the attributes of those used for training, aiming to evaluate the quality of feature matching performance and camera pose estimates in synthetic noise. In addition, we also test on real-world drone bursts captured in millilux illumination for 3D reconstruction.

5.4 Results

5.4.1 Implementation Details

We aim to improve feature detection and extraction in robotic bursts, enabling improved performance in low-light 3D reconstruction. LBurst, our learning-based robotic burst feature finder, operates on a burst of images at different scales to compute sparse locations that correspond to keypoints with well-defined features and their corresponding descriptors. During inference, LBurst analyzes the burst at multiple scales, and the features are then scored based on their detection and description confidence score. To evaluate the capabilities of LBurst, we conduct synthetic experiments using the generated HPatches burst dataset. We further demonstrate 3D reconstruction of real-world scenes in millilux conditions along multiple drone trajectories, utilizing actual bursts captured by drones.

Training Datasets

We train our method using two sources of single image-based datasets. We incorporate 4,479 images from the Aachen dataset [84], which captures scenes resembling those captured by drones in urban cityscapes and 3,190 random web images [147] as original images in constructing 5-frames robotic bursts. We exclude captured robotic bursts during training and demonstrate our model generalizing well towards drone-captured data.

Training Parameters

We generate a 5-frames robotic burst for each image by incorporating image translations and noise following a normal distribution. The maximum translation between frames is limited to 30 pixels and the noise level variance of the images falls within the range of 0.3 to 0.6 on normalized images. We perform random homographic transformations between bursts during training to improve feature extraction. For network optimization, we employ the Adam optimizer [148]. The training process involves 25 epochs with a fixed learning rate of 0.0001. We also apply weight decay of 0.0005 to regularize the model.

5.4.2 Feature Performance in Noise

We compare our approach of adapting R2D2 for a robotic burst with the original 2D R2D2 on images that exhibit significant noise to evaluate the feature performance in synthetic noise. This is because an end-to-end approach for single-image feature extraction has the potential to learn how to interpret features in noisy images when provided with sufficient training data, to an extent. Additionally, the impact of temporal information in enhancing the learning-based feature extraction on low SNR images remains largely unexplored.

We generate bursts from the HPatches dataset [128], which consists of a sequence of images with increasing difficulty in terms of viewpoint and illumination. We use

each image in the dataset as an original image to generate a robotic burst with uniform translation and random noise as described in [Section 5.3.2](#) and by doing so, we create bursts that exhibit varying viewpoints and illumination for each scene. For camera pose estimation, we only use bursts generated with varying viewpoints between images, i.e., 56 HPatches dataset scenes. We further compare against existing state-of-the-art classical and learning-based techniques on conventional images, as well as other robotic burst imaging methods discussed in [Section 5.4.4](#).

Matching Accuracy in Synthetic Noise

Matching accuracy demonstrates the ability of a model to correctly match pixel-level correspondences between the same scene. High matching accuracy ensures accurate 3D reconstruction allowing improved correspondence search. In contrast, poor matching accuracy leads to incorrect or incomplete reconstructions. We compare the mean matching accuracy (MMA) for various pixel error thresholds using our method on light-constrained bursts and state-of-the-art [R2D2](#) method on corresponding common images for all 108 scenes. We compute ground truth by operating [R2D2](#) on

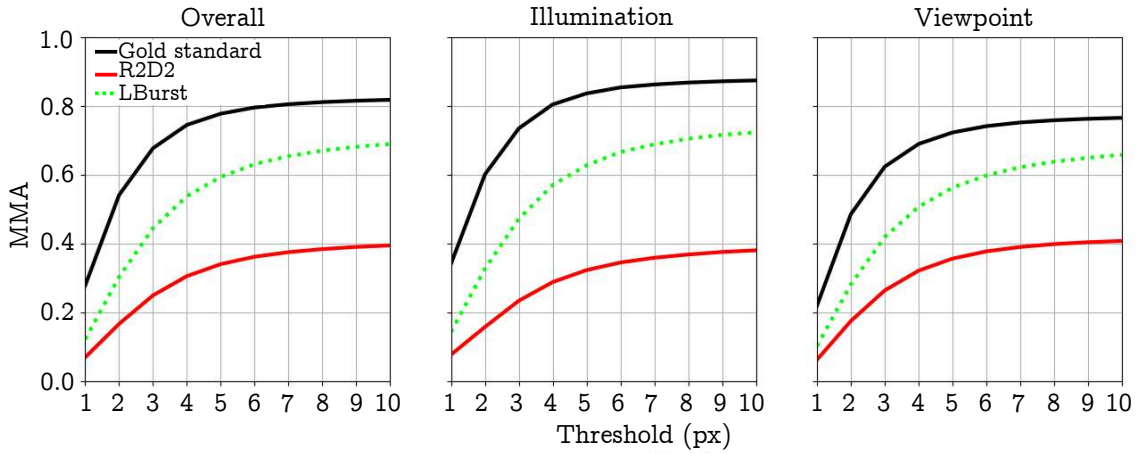


Figure 5.6 – Feature matches on synthetic burst and corresponding images for the HPatches dataset. Overall, our burst method outperforms [R2D2](#) learned feature extraction (left) with improved illumination (middle) and viewpoint (right) for light-constrained images. This demonstrates the viability of our approach for feature-based reconstruction using light-constrained scenes.

original images in the HPatches dataset. Our method shows improved matching performance compared to conventional methods for all threshold values as in Figure 5.6 in challenging conditions for reconstruction performance.

We evaluate mean matching accuracy, as defined in [43], the percentage of true matches between a pair of bursts with an error threshold of 3 pixels and repeatability score by calculating the number of putative matches computed for detected keypoints as shown in Table 5.1. Our method shows two times the matching accuracy of R2D2 operating on noisy images and demonstrates improved repeatability score. In the table, green bold highlights the overall best results.

By adjusting the level of noise introduced during inference on the generated HPatches bursts, we assess the role of noise in matching performance. We compare against R2D2 on the original common image (black) and our method on the generated robotic bursts. We demonstrate that our method outperforms R2D2 in both moderate and strong noise while showing a decline in matching performance with increasing noise as shown in Figure 5.7.

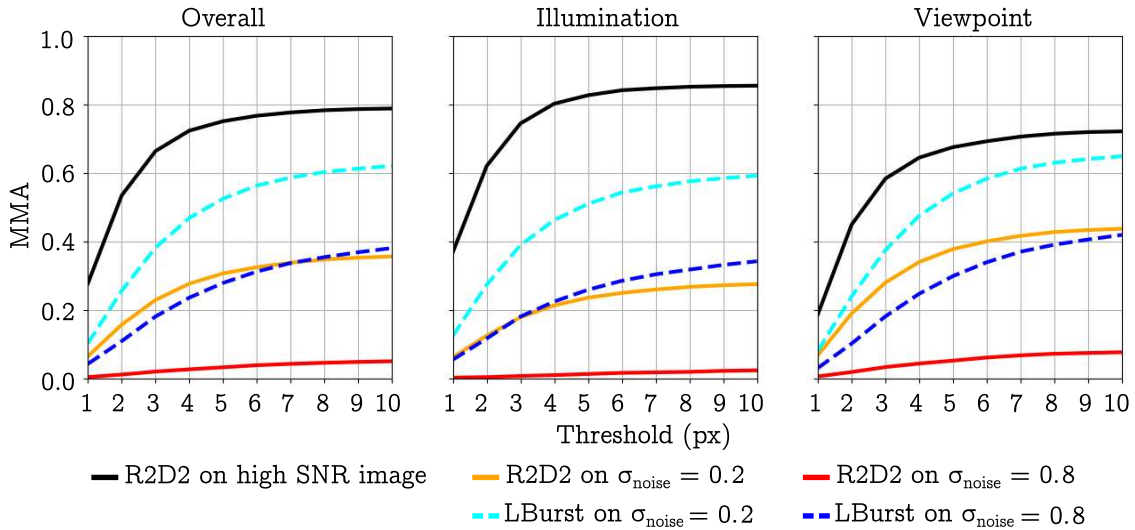


Figure 5.7 – We operate R2D2 on the original HPatches dataset with high SNR (black). We introduce noise to mimic light-constrained scenes and demonstrate R2D2 fails to produce accurate matches for moderate (orange) and strong (red) noise while our method outperforms state-of-the-art algorithms in producing improved feature matches in both moderate (cyan) and strong (blue) noise.

Table 5.1 – Mean matching accuracy and repeatability score for all images in the generated HPatches dataset: Our learned robotic burst feature finder approach outperforms R2D2 2D feature extraction on the common image of the corresponding robotic burst. Bold, Green: Overall best results from our proposed approach.

Method	MMA@3	Repeatability
Gold Standard	0.68	0.31
R2D2	0.25	0.10
LBurst (Ours)	0.44	0.29

Camera Trajectory Accuracy in Synthetic Noise

We select a subset of the HPatches dataset with varying viewpoints to evaluate camera poses. We show an example scene (viewpoint - vitro) and the corresponding camera pose estimates in Figure 5.8. We use R2D2 on original images (black) to collect gold standard poses using COLMAP. We compare our method, which operates on generated HPatches bursts (green), against R2D2, which operates on the common images (red). We align and scale the estimated camera poses to the gold standard poses using Sim(3) Umeyama’s alignment method [119] similar to Chapter 3 and Chapter 4 as there is an arbitrary scale factor involved in monocular SfM.

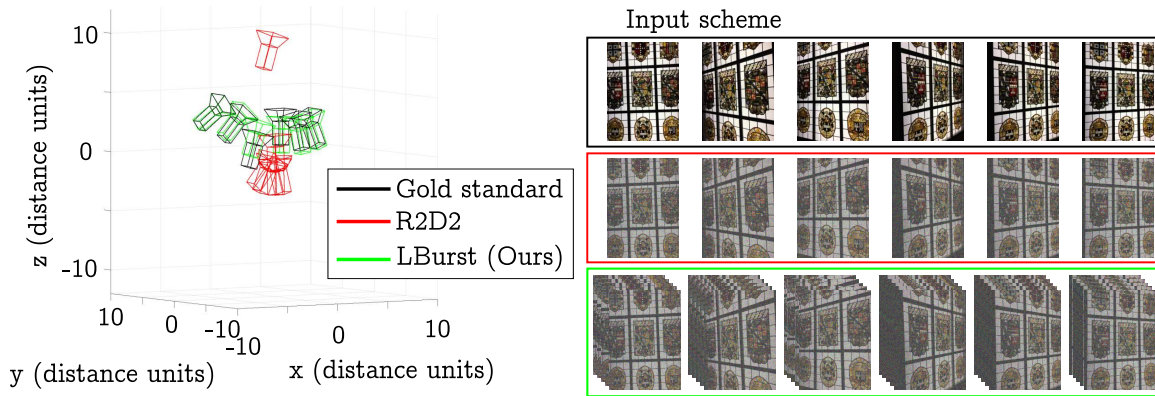


Figure 5.8 – (Right) (top row): A scene from HPatches original dataset (viewpoint - vitro), which poses a viewpoint challenge for feature detection; (bottom row): A generated burst of images from each original input; (middle row) common image of the generated burst used for conventional 2D feature extraction. (Left) Estimated camera trajectory using COLMAP. Our method (green) reconstructs more accurate camera poses compared to R2D2 (red) and correspond closely to the gold standard poses from original high SNR images (black).

Table 5.2 – Mean translation error and mean rotational error in camera poses for reconstruction methods of a subset of the generated HPatches dataset. The results are in distance units. Bold, Green: Overall best results from our proposed approaches.

Method	Absolute trajectory error		Relative pose error	
	trans.	rot.	trans.	rot.
R2D2	3.30	2.18	5.18	0.86
LBurst (Ours)	1.32	1.03	1.90	0.45

We additionally compute average absolute instantaneous and relative pose errors (translation and rotation) against gold standard poses, as in Table 5.2. Our method converges for 60% of diverse viewpoint scenes, outperforming R2D2 (12.5%) in low-light image scenarios with accurate pose estimates on the generated HPatches dataset. The color scheme matches that used in Table 5.1. The bold green values indicate that our method outperforms alternative approaches across all metrics. This shows the feasibility of using learned robotic burst feature finder as an input scheme for correspondence search in light-constrained reconstruction.

5.4.3 Reconstruction Performance

We evaluate reconstruction performance using our previous dataset captured in light-constrained conditions as described in Chapter 4 in addition to drone-based validation in the following section. This is because the drone imagery captures scenes far away demonstrating a small amount of apparent motion while some other robotic platforms and trajectories have the tendency to produce more parallax motion.

Table 5.3 shows reconstruction performance using our method compared to state-of-the-art feature extraction methods for reconstruction. In the table, green highlights the best results from the proposed method and red shows the best results from competing approaches. We show overall improved reconstruction performance with improved inlier matches, match score and more 3D points compared to other robotic burst imaging methods resulting in a more complete reconstruction.

Table 5.3 – Average performance of reconstruction for light-constrained scenes with 1D and 2D apparent motion including learned robotic burst feature finder. Bold: Overall best results, Green: Best results from our proposed approaches; Red: Best results from alternative approaches.

Apparent Motion	Method	Convergence Rate	Images Pass %	Keypoints/ Image	Putative Matches/ Image	Inliers Matches/ Image	Match Ratio	Match Score	Precision	3D Points/ Image
1D	Gold Standard	1.0	100	2880.96	1.18x10 ⁴	1.17x10 ⁴	4.16	4.14	0.99	257.40
	SIFT	0.0	0	5573.68	6.84x10 ¹	6.41x10 ¹	0.01	0.01	0.87	0.00
	Conventional Noisy	0.0	0	29.48	4.50x10 ⁻¹	0	0.02	0.00	0.00	0.00
	R2D2	0.0	0	2000.00	1.32x10 ³	1.72x10 ²	0.66	0.09	0.13	0.00
	SIFT	0.8	59	2216.14	3.62x10 ²	3.32x10 ²	0.19	0.17	0.93	31.60
	Burst with Merge	1.0	69	154.64	1.26x10 ¹	7.45x10 ⁰	0.58	0.29	0.56	4.20
	R2D2	0.6	37	2000.0	1.20x10 ³	2.18x10 ²	0.60	0.11	0.18	3.20
	BuFF 1D	1.0	98	1386.30	3.84x10 ²	3.75x10 ²	0.28	0.27	0.98	36.50
	BuFF 2D	1.0	92	1635.82	3.45x10 ²	3.38x10 ²	0.23	0.23	0.98	36.20
	LBurst (Ours)	1.0	100	2751.21	1.42x10³	1.39x10³	0.52	0.50	0.98	73.90
	Gold Standard	1.0	100	3270.16	1.30x10 ⁴	1.29x10 ⁴	3.89	3.87	0.99	305.50
	SIFT	0.0	0	5680.10	1.84x10 ¹	1.43x10 ¹	0.00	0.00	0.86	0.00
2D	Conventional Noisy	0.0	0	25.88	3.38x10 ⁻¹	0	0.01	0.00	0.00	0.00
	R2D2	0.0	0	2000.00	1.30x10 ³	1.88x10 ²	0.65	0.10	0.15	0.00
	SIFT	0.8	77	2745.74	4.85x10 ²	4.57x10 ²	0.18	0.17	0.95	32.20
	Burst with Merge	0.4	19	94.40	2.87x10 ¹	8.15x10 ⁰	0.24	0.07	0.24	10.84
	R2D2	0.4	41	2000.00	1.25x10 ³	1.95x10 ²	0.62	0.10	0.16	3.99
	BuFF 1D	0.8	85	2572.40	5.55x10 ²	5.43x10 ²	0.21	0.21	0.95	42.20
	BuFF 2D	1.0	98	2143.96	5.65x10 ²	5.51x10 ²	0.26	0.26	0.98	47.20
	LBurst (Ours)	1.0	100	2694.67	1.38x10³	1.15x10³	0.44	0.43	0.97	73.10
	Gold Standard	1.0	100	3270.16	1.30x10 ⁴	1.29x10 ⁴	3.89	3.87	0.99	305.50
	SIFT	0.0	0	5680.10	1.84x10 ¹	1.43x10 ¹	0.00	0.00	0.86	0.00
	Conventional Noisy	0.0	0	25.88	3.38x10 ⁻¹	0	0.01	0.00	0.00	0.00
	R2D2	0.0	0	2000.00	1.30x10 ³	1.88x10 ²	0.65	0.10	0.15	0.00

Table 5.4 – Mean translation error and mean rotational error in camera poses for reconstruction methods of light-constrained scenes with 1D and 2D apparent motion, including learned robotic burst feature finder. Bold: Overall best results, Green: Best results from our proposed approaches; Red: Best results from alternative approaches.

Apparent motion	Method	Absolute trajectory error		Relative pose error	
		trans.(cm)	rot.(deg)	trans.(cm)	rot.(deg)
1D	Burst with Merge - SIFT	2.84	2.07	4.97	0.73
	Burst with Merge - SuperPoint	2.13	2.17	5.12	0.81
	Burst with Merge - R2D2	1.71	1.20	3.17	0.04
	BuFF 1D	1.23	1.24	2.00	0.05
	BuFF 2D	1.27	0.62	2.01	0.02
	LBurst	1.64	1.04	2.33	0.05
2D	Burst with Merge - SIFT	1.80	1.11	3.60	0.04
	Burst with Merge - SuperPoint	1.56	1.06	3.71	0.04
	Burst with Merge - R2D2	1.48	0.55	2.84	0.04
	BuFF 1D	1.53	1.17	2.99	0.05
	BuFF 2D	1.44	0.58	2.55	0.03
	LBurst	1.35	1.03	2.63	0.04

In [Table 5.4](#), we present a comparison of our approach for estimating camera poses with other methods. Our learned robotic burst feature estimates outperform other methods in estimating absolute trajectory error in translation while the variants of robotic burst feature finder perform well for most cases.

5.4.4 Drone-based Validation

We validate learned burst features for low-light 3D reconstruction by deploying drones, a DJI Mini Pro 3 and a DJI Phantom Pro 4 in millilux illumination as shown in [Figure 5.9](#). We record multiple trajectories during drone flights, comprising 10 scenes with 15 bursts each. Each burst consists of 5 images and exhibits approximately linear 2D motion within each burst.

Experiment Details

During drone flights, we capture multiple trajectories, with each trajectory consisting of 15 bursts, and each burst containing 5 images. These robotic burst sequences

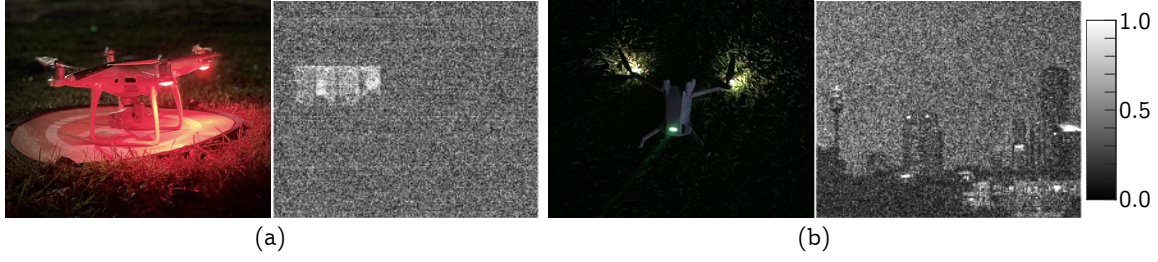


Figure 5.9 – We use two drones to capture light-constrained scenes. (a) (left) DJI Phantom Pro 4 captures images over the exposure of 0.125 ms at millilux conditions. (right) An example image in a captured burst which demonstrates noise-limited nature. (b) (left) DJI Mini Pro 3 captures images over the exposure and millilux conditions. (right) An example image in a captured burst of drone imagery. The full dataset contains 10 scenes, each with 15 bursts of 5 images. (Contrast-adjusted image for visualisation)

exhibit approximately linear 2D motion within each burst. We use two different drones, DJI Mini Pro 3 and DJI Phantom Pro 4, both equipped with monocular cameras. The DJI Mini Pro 3 camera has an $f/1.7$ lens, captures 16-bit raw images of size 4032×2268 over a duration of 0.125 seconds, and uses an ISO setting of 6400. On the other hand, the DJI Phantom Pro 4 camera has an $f/2.8$ lens, captures 16-bit images of size 4864×3648 over the same duration of 0.125 seconds, and uses an ISO setting of 12800. During the data collection stage, the lighting conditions varied in millilux illumination as measured using a light meter. While we set both cameras to maintain the same EV across platforms, due to reduced natural light in extreme low-light conditions, the imagery captured using the DJI Phantom Pro 4 drone exhibited stronger noise compared to that captured using the DJI Mini Pro 3 drone.

Feature Performance on Drone Imagery

We show two light-constrained scenes captured by DJI Phantom Pro 4 (top) and DJI Mini Pro 3 (bottom) in Figure 5.10. In particular, we demonstrate the features detected by LBurst, along with an interpretable confidence map for both detection and descriptor. The majority of the features detected by our method are highly robust, avoiding unreliable matching regions such as the sky under millilux conditions. This is

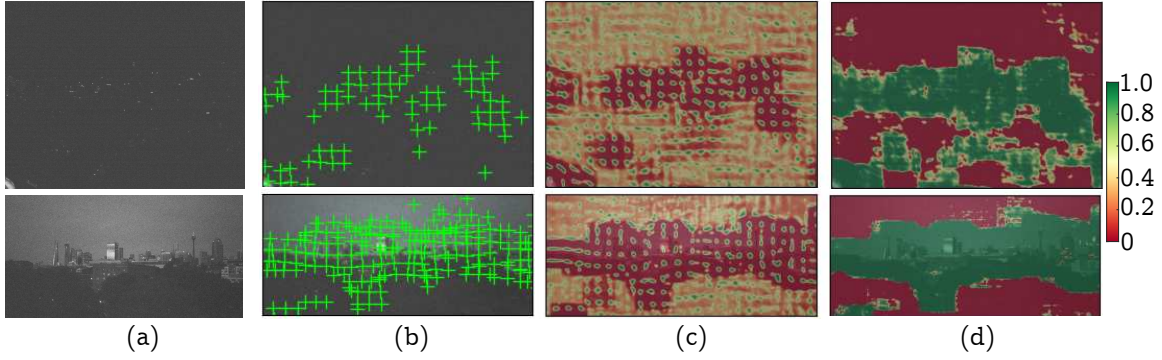


Figure 5.10 – We locate keypoints using LBurst on drone images captured using two commercial drones. (top row) Light-constrained scenes with strong noise captured by a DJI Phantom Pro 4 at millilux conditions of 0.02 - 0.08 lux. (bottom row) Light-constrained scenes with moderate noise captured by a DJI Mini Pro 3 at millilux conditions of 0.05 - 0.12 lux. For both drones, we present (b) the top scored keypoints (c) detection confidence map, and (c) descriptor confidence maps.

because we evaluate the repeatability and reliability scores of every detected feature, selecting the features with high scores for reconstruction.

We compare our method (top) to the state-of-the-art [R2D2](#) (bottom) on a robotic burst and demonstrate a qualitative improvement in detected keypoints and their corresponding maps with our method compared 2D feature detection as in [Figure 5.11](#).

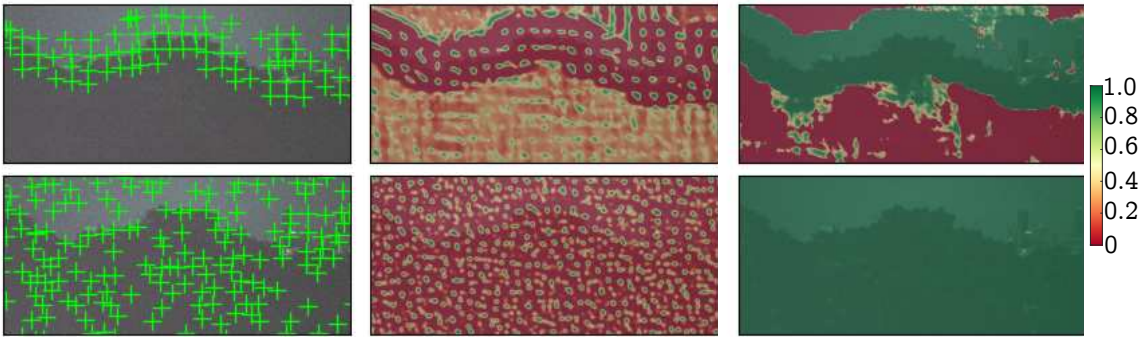


Figure 5.11 – We compare (top row) our method LBurst against (bottom row) state-of-the-art [R2D2](#) on a light-constrained scene. (left) LBurst demonstrates high-quality true features and fewer spurious features compared to [R2D2](#). (middle) Our detection confidence maps show repeatable regions as local extrema while [R2D2](#) fails to produce repeatable regions. (right). Our descriptor confidence map shows the reliable region avoiding sky regions for reconstruction in contrast to [R2D2](#) finding all regions as reliable.

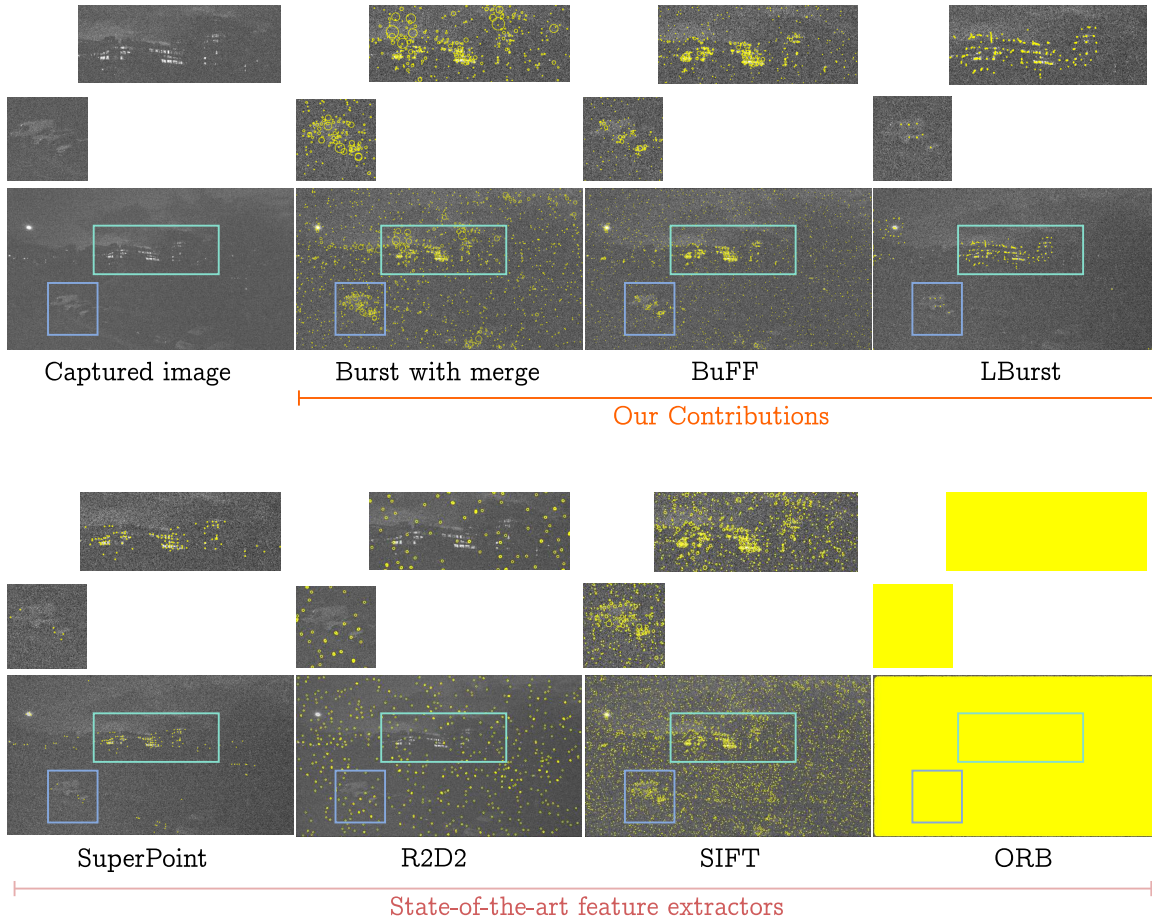


Figure 5.12 – Comparison of feature performance on a captured scene (top, left) using conventional burst techniques (top row) and single image based techniques (bottom row), highlighting features associated with background (green) and foreground (blue) regions. Burst methods outperform conventional feature extraction methods, and learned burst feature finder outperforms other burst methods by generating more robust features of the scenes.

This demonstrates how our method benefits from the temporal information in a burst to handle strong noise and small motion variation on captured burst imagery.

Finally, we evaluate the performance of our feature detection method against state-of-the-art classical and learning-based methods on conventional images and other robotic burst imaging methods as shown in Figure 5.12. For a captured scene using DJI Mini Pro 3, we demonstrate burst-based methods locating most of the high-quality true features corresponding to the vehicle in the foreground (blue) and the buildings in the background (green) with BuFF showing slightly higher performance in rejecting

spurious features compared to burst with merged image using [SIFT](#). Our learned robotic burst feature finder avoids most of the spurious features and low contrast regions and locates both the vehicle and buildings. It also finds the light source which has higher pixel intensity, and so local extrema within bursts.

We also compare our method with learning-based methods like SuperPoint and [R2D2](#), as well as classical methods such as [SIFT](#) and [ORB](#) at default settings. Although SuperPoint exhibits strong performance, our approach generates twice as many features as SuperPoint, leading to a higher number of inlier matches during reconstruction. [R2D2](#) performs poorly compared to other learning methods. [SIFT](#) struggles to find high-quality true features and produce more spurious features compared to learning and burst-based methods. [ORB](#) locates overwhelming features failing to represent this light-constrained scene for reconstruction.

Reconstruction Performance on Drone Imagery

Following the feature comparison approach in [\[137\]](#), we evaluate reconstruction performance for the drone imagery similar to [Chapter 3](#) and [Chapter 4](#). We compute a maximum of 4000 keypoints with thresholds on detection and descriptor at 0.7 for both our approach and [R2D2](#). We evaluate against other methods, including the conventional single-image capture and burst-merged image approaches ([Chapter 3](#)), which utilize classical method such as [SIFT](#) and learning method including [R2D2](#) and SuperPoint, as well as the robotic burst feature finder ([Chapter 4](#)), as shown in [Table 5.5](#). The color scheme matches that used in [Table 5.3](#).

Our method outperforms state-of-the-art feature detection methods on both conventional images and burst-merged methods, by reconstructing the majority of scenes captured by the DJI Mini Pro 3 with a higher number of 3D points. In comparison, our approach achieves twice as many inlier matches as the physics-based burst feature finder method. For images captured by the DJI Phantom Pro 4, our method reconstructs all scenes with the highest number of inlier matches and 3D points, resulting in more complete 3D sparse models. While SuperPoint takes advantage of

Table 5.5 – Average performance of reconstruction for light-constrained drone imagery. Bold: Overall best results, Green: Best results from our proposed approaches; Red: Best results from alternative approaches.

	Methods	Convergence rate	Images passed %	Keypoints/ Image	Putative matches/ Image	Inliers/ Image	Match ratio	Match score	Precision	3DPoints/ Image
DJI Mini Pro 3	Conventional									
			14.70	8202.16	83.96	71.93	0.025	0.021	0.518	10.46
			30.70	936.10	610.30	208.10	0.599	0.296	0.439	2.59
	R2D2	0.20	20.00	4000.00	1347.74	310.40	0.337	0.008	0.235	1.63
	Burst with Merge									
			45.30	6628.68	197.51	169.41	0.061	0.052	0.805	23.06
		0.80	72.00	1044.96	728.38	427.90	0.652	0.371	0.508	20.35
	R2D2	0.80	76.00	4000.00	546.30	320.90	0.137	0.080	0.508	9.86
	BuFF	0.80	64.00	5603.19	177.72	158.43	0.084	0.074	0.814	25.41
	LBurst (Ours)	0.80	81.30	4000.00	690.82	351.16	0.173	0.088	0.508	31.40
DJI Phantom Pro 4	Conventional									
			0.00	6116.08	5.71	2.44	0.002	0.002	0.169	0.00
		0.40	40.00	533.28	393.28	230.50	0.701	0.386	0.523	5.13
	R2D2	0.00	0.00	4000.00	1025.57	103.52	0.256	0.026	0.101	0.00
	Burst with Merge									
			0.00	2218.68	27.91	17.80	0.013	0.008	0.580	0.00
		0.40	40.00	495.84	362.28	219.20	0.704	0.406	0.564	6.35
	R2D2	0.40	40.00	4000.00	1346.90	420.68	0.337	0.105	0.300	5.40
	BuFF	0.40	28.00	1483.42	17.74	14.71	0.032	0.027	0.859	11.78
	LBurst (Ours)	1.00	89.30	4000.00	1179.38	480.42	0.295	0.120	0.432	14.85

the SNR boost associated with burst-merged images and demonstrates competitive performance in detecting inlier matches, it struggles to identify 3D points. On the other hand, R2D2 is designed to optimize AP, similar to our approach, but it tends to generate an excessive number of spurious putative matches. Overall, our method benefits from improved feature performance, which leads to a more complete 3D reconstruction in low light with the highest convergence rate.

We qualitatively analyze light-constrained reconstruction in Figure 5.13. We select a trajectory captured as multiple bursts by DJI Mini Pro 3 in low light and for which most methods reconstruct. State-of-the-art feature extractors, both classical and learning-based methods demonstrate poor performance on conventional images. This is because the amount of signal in captured imagery is low for feature extractors to find high-quality features. In addition, they also produce spurious features and low quality matches for 3D reconstruction. We operate SIFT feature extraction on temporally-merged images as discussed in Chapter 3 and also operate SuperPoint and R2D2 with a burst with merge approach. We demonstrate improved reconstruction performance with the signal boost from the burst-with-merge approach.

We apply a physics-based robotic burst feature finder to find well-defined blob features directly in a burst as in Chapter 4 and demonstrate improved feature performance compared to conventional and burst-merged methods. Overall, as highlighted, the learned robotic burst feature finder reconstructs the scene with all input images, more 3D points, and a more complete structure compared to alternative approaches.

5.4.5 Speed

We repeatedly run the network by varying scales to compute scale-invariant features for reconstruction. We keep the largest dimension to 2^{10} pixels and downsample to 2^8 at the steps of $2^{1/4}$. We locate features by finding the local maxima at each scale with a patch size of $N = 16$. We score the descriptors as a product of repeatability and reliability and select the best-scored descriptors for matching. In practice, it takes an average of 2.27 seconds on a NVIDIA GeForce RTX 3080 Laptop GPU to extract

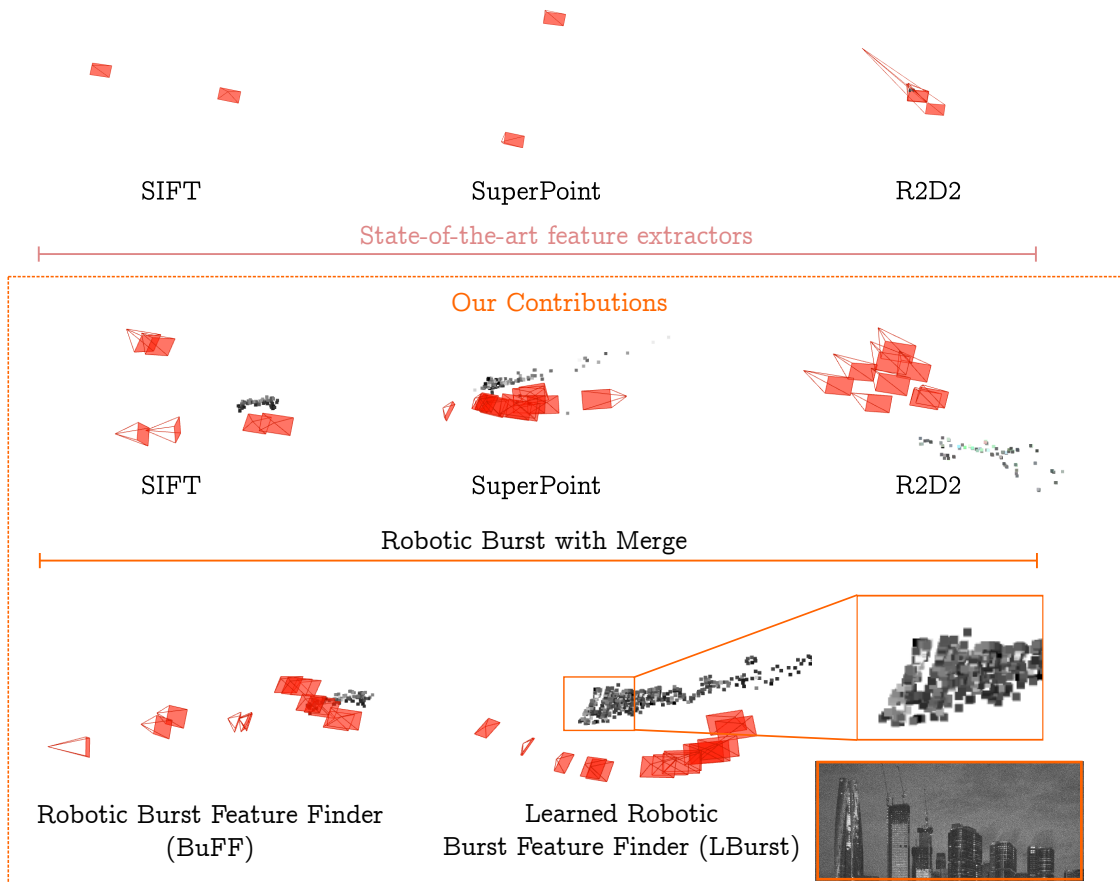


Figure 5.13 – Overall reconstruction performance for a light-constrained scene (bottom right). (Top row): Existing state-of-the-art feature extractors including [SIFT](#), SuperPoint and [R2D2](#) fail to perform reconstruction using conventional images; (middle row): We operate SIFT, SuperPoint and R2D2 on burst-merged images which demonstrate progressive improvement in reconstruction compared to conventional approaches. We employ robotic burst feature finder (bottom, left) and learned robotic burst feature finder (bottom, right) on drone imagery and demonstrate more registered images and higher 3D points resulting in more complete reconstruction compared to conventional methods. Overall, feature finder outperforms state-of-the-art feature extraction techniques and alternative robotic burst imaging methods.

keypoints at all scales on a burst of 5 drone images and 1.34 minutes for a complete reconstruction using COLMAP.

5.4.6 Ablation Study

We explore the variables in the learned robotic burst feature finder for visual reconstruction.

Keypoints. We change the number of keypoints computed by our model as shown in Figure 5.14. Our results indicate that the top-scoring detection exhibits keypoints that directly improve reconstruction performance, producing higher-quality matches. This is useful for applications with varying needs from sparse to dense pixel-level correspondences. However, beyond a certain point, the inclusion of more keypoints may introduce spurious features.

Patch size. We change the patch size during training and inference, which affects the number of features during multi-scale feature detection as shown in Figure 5.15. Increasing the patch size corresponds to sparsely located features of an image compared to a small patch size. This is because the features are detected which correspond to the local extrema of a patch operating on all image pixels. However, there is a trade-off between the number of detected points and their confidence score. While a smaller patch size yields a denser distribution of keypoints and increases the number of detected features, it can introduce grid-like patterns and reduce robustness, especially in low SNR images. This could become the bottleneck of our approach, which can be addressed by adaptively selecting the patch size or by explicitly modeling the motion behaviours within a burst.

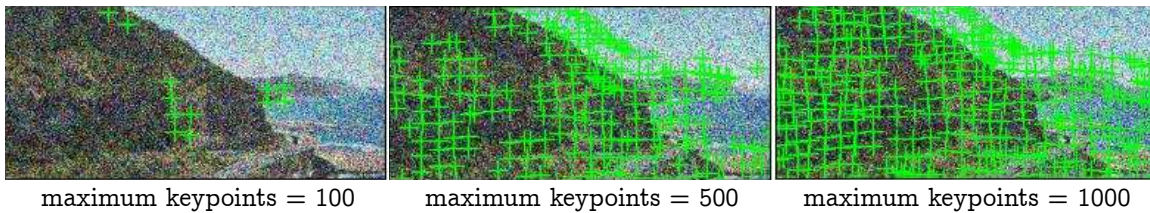


Figure 5.14 – We demonstrate associated keypoints for the same burst of images by increasing the number of keypoints detected in a burst. This shows the feasibility of using our method for various applications that require sparse and dense keypoints.

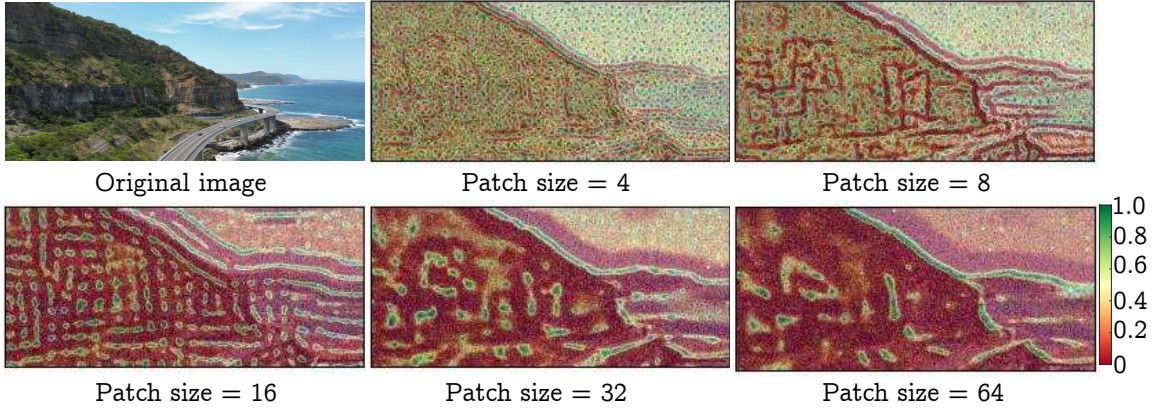


Figure 5.15 – We demonstrate the impact of patch size in our model using a robotic burst of the high SNR image (top row, left). Models trained on different patch sizes yield varied detection maps. When the patch size is small, repeatable features (green) appear densely on the sky in a grid pattern (top-middle, right), detecting non-salient keypoints. Increasing the patch size (bottom row) allows for sparser feature detection by searching for local extrema in larger patches.

5.4.7 Failure Cases

Our method is prone to false reliable scores for descriptors in the presence of strong noise as shown in Figure 5.16. This is because, similar to R2D2 and L2-Net, we optimize for descriptor loss by maximizing the AP for a robotic burst. As we increase the noise, the descriptor confidence remains high, while the detection confidence map demonstrates spurious detected features, such as a grid pattern in the sky region. When we compute keypoints from high-scored, detection and description confidence maps, this introduces spurious features in situations where no reliable features are present in the scene.

An example of a failed case due to strong noise and low contrast is shown in Figure 5.17. Although feature detection using our method (top) performs better in strong noise compared to R2D2 (bottom), the reliability information is still limited. Our approach produces spurious putative matches, like those from river regions, which may not be beneficial for reconstruction. Additionally, it yields fewer inlier matches in lower contrast regions for light-constrained reconstruction. We believe capturing a burst with more images will benefit feature detection and description.

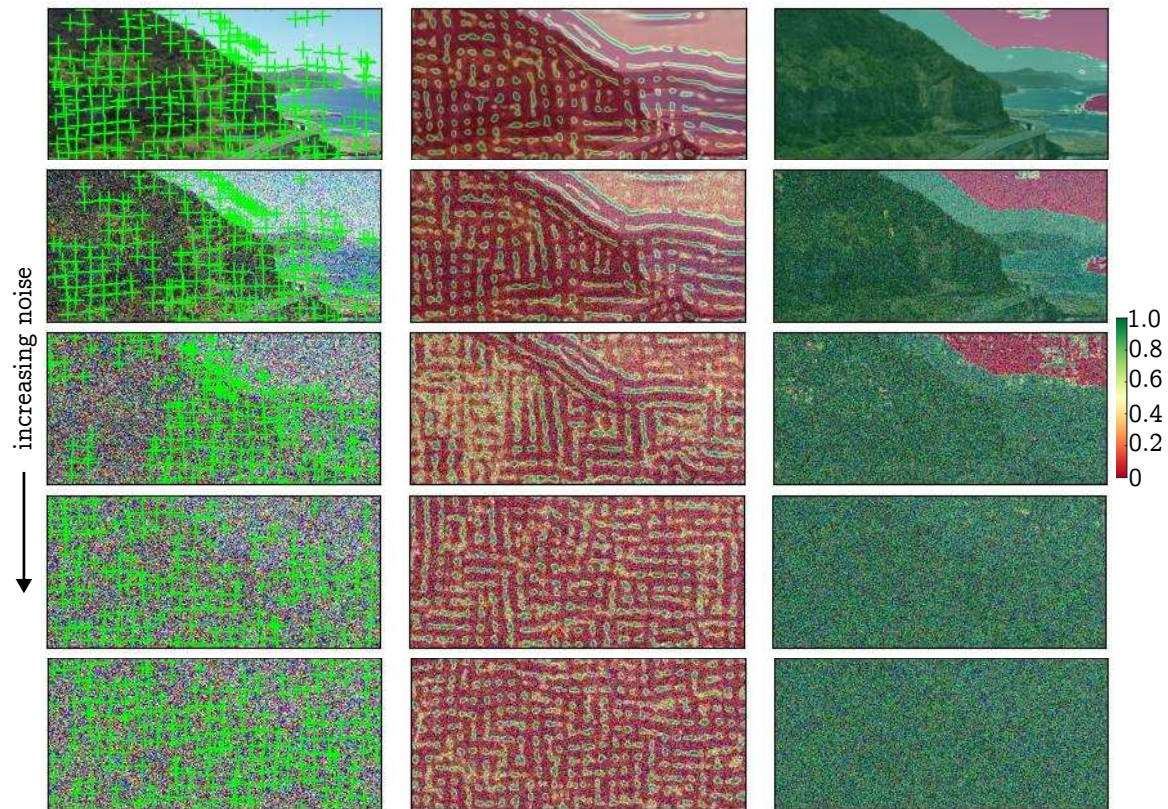


Figure 5.16 – Our method starts detecting spurious features (left) with low repeatable features (middle) and detecting all regions as reliable (right) as we increase the noise within the burst images. This demonstrates the limitation of our current model with strong noise.

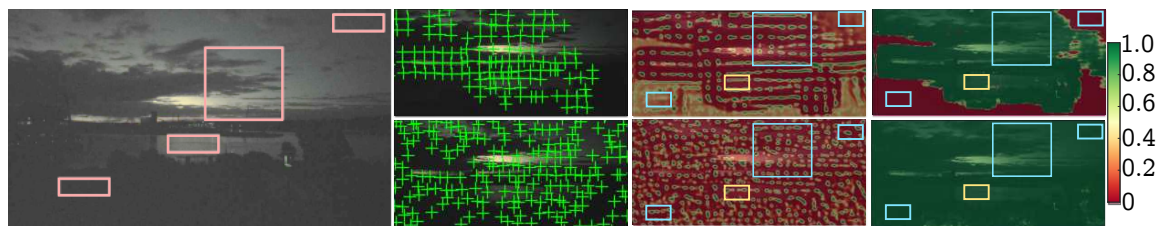


Figure 5.17 – We show a detailed failure case example for our method (top) and compare it against state-of-the-art [R2D2](#) (bottom). (left) A light-constrained scene captured at twilight (contrast adjusted for visualisation). We demonstrate four interesting areas in the image - low contrast region, a river, and two sky regions with and without a cloud pattern. We demonstrate in three cases (blue) our method outperforms [R2D2](#) and fails to identify pixels corresponding to the waterfront as non-reliable regions (yellow).

5.5 Discussion and Future Directions

We developed a feature finder based on learned robotic bursts for 3D reconstruction of drone imagery in extremely low-light conditions. Our method detects and describes features directly from light-constrained robotic bursts, employing unsupervised learning and metric learning loss to locate well-defined noise-tolerant features. By allowing our model to map directly from a burst with temporal information, we demonstrated overall improved reconstruction performance in millilux conditions.

We exhibited improved matching accuracy and camera pose estimations on the generated HPatches dataset over a conventional learning-based feature detector, [R2D2](#), on robotic bursts. We validated reconstruction performance using BuFF dataset and drone imagery captured at millilux illumination. Our method showed overall improved performance compared to other robotic burst imaging approaches including burst with merge as described in [Chapter 3](#) and robotic burst feature finder [Chapter 4](#). We also compared our method to conventional methods and learning-based feature extraction on conventional images, and our approach demonstrated complete models with higher 3D points and convergence rates in light-constrained scenes.

We also showed improved reconstruction performance and camera pose estimation on the [BuFF](#) dataset with larger 1D and 2D apparent motion compared to drone burst imagery. While the [BuFF](#) variants provided strong camera pose estimates for this dataset against the learned robotic burst feature finder, we believe that fine-tuning our training model on a compact dataset, which presents significant parallax movement between frames, could further optimize pose estimation accuracy.

We observed limited reliability of our method at low contrast and strong noise, leading to more putative matches. We believe training in a multi-scale space, rather than only computing multi-scale features during inference, might improve feature performance in low-light conditions. In addition, we anticipate that incorporating complementary sensors could enhance our burst-based reconstruction approach, allowing robots to operate even in challenging millilux illumination conditions.

Chapter 6

Conclusions and Future Directions

"Somewhere, something incredible is waiting to be known"

- Carl Sagan

6.1 Conclusions

The field of 3D reconstruction has witnessed significant advancements in recent years. However, state-of-the-art reconstruction techniques [20, 39] had limitations when operating in low-light conditions due to the low SNR of captured images. In this thesis, we identified an opportunity to advance 3D reconstruction in low light using conventional cameras by adapting burst imaging for robotic vision. Observing the success of feature-based methods for reconstruction with large viewpoint and intensity variations, we proposed to employ robotic burst imaging for feature-based reconstruction to represent sparse images that would otherwise have been captured in a conventional manner and by doing so, extended the low light in which robots operate.

We adapted burst photography, commonly used in mobile photography [33] for reconstruction in low light by enabling the use of direct methods for image registration within the burst of images and using feature-based methods to handle the sparsity between bursts of images for reconstruction in Chapter 3. In this contribution,

we demonstrated successful reconstruction of light-constrained scenes with decreased failure cases due to non-convergence compared to alternative imaging methods of capturing a single image or an image sequence as an input scheme for correspondence search. We showed improved feature performance and camera pose estimates and established robotic burst imaging as a successful candidate for light-constrained reconstruction. To the author’s knowledge, this work is the first step toward solving the problem of 3D reconstruction in low light using robotic burst imaging.

Taking a step further to deal with misalignment at strong noise when performing reconstruction in extremely low light, we introduced a novel robotic burst feature finder to reconstruct low-light scenes in [Chapter 4](#). We directly found blob features with well-defined scale and apparent motion within bursts of images by searching for local extrema through scale-slope space and employed a descriptor that provided additional advantage with higher SNR boost. By directly finding well-defined features in a higher dimensional search space we demonstrated high-quality true features and lower spurious features compared to state-of-the-art feature extractors on conventional images and burst-merged images. We also demonstrated improved reconstruction performance with 3D detection over conventional 2D detection on burst imagery as described in our previous contribution [Chapter 3](#).

With the success of directly finding features within a burst of images as described in [Chapter 4](#) over 2D detection on bursts of images as discussed in [Chapter 3](#), we extended 2D learning-based feature detection to operate on a robotic burst. By doing so, we improved the accuracy of feature matching and surpassed the limitations imposed by physics-based approaches. We computed burst features that are noise-tolerant by employing an unsupervised loss to deal with similarity between bursts during training and leveraging metric learning loss to compute corresponding descriptors in reliable regions. We evaluated our learned burst feature finder against state-of-the-art feature extractors on conventional images and burst-merged images. We also evaluated against physics-based burst features as discussed in [Chapter 4](#) and demonstrated overall improved feature performance and competitive 3D reconstruction performance.

We validated our approach by capturing datasets using a robotic arm in controlled light settings for [Chapter 3](#) and [Chapter 4](#) changing the low-light conditions and cameras to progressively introducing challenging image scenes for each contribution. We also captured outdoor light-constrained scenes at night under millilux illumination conditions using DJI Phantom Pro 4 and DJI Mini Pro 3 drones as introduced in [Chapter 5](#). We evaluated state-of-the-art feature extraction techniques and all of our robotic burst imaging techniques as input methods in searching for correspondences to reconstruct scenes under low-light conditions of millilux intensity.

In this thesis, we demonstrated an ability to operate robots in lower light than was previously possible using a robotic burst imaging scheme. We believe that this work opens the possibilities for a broad range of applications including autonomous driving and delivery drones at night, mining, and understanding nocturnal behaviors of animals. We anticipate future directions to extend the lower light in which robots currently operate and have described few ways of doing so in the following section.

6.2 Future Directions

We allow robots to operate in low light, by advancing light-constrained 3D reconstruction using robotic burst imaging. In particular, we demonstrate adapting burst imaging for robotic vision on feature-based [SfM](#). An immediate application-driven future direction is to use the developed methods for current open research problems at night such as object detection, feature-tracking and low-light neural radiance fields ([NeRF](#)). In the following, we consider a few long-term research avenues for robotic burst imaging.

6.2.1 Adaptive Robotic Burst Imaging

There are a few ways of performing adaptive robotic burst imaging for reconstruction. In the following, we consider two methods of doing so to achieve cheap computation and improved performance.

Active vision. Robotic burst imaging benefits from large overlap between images in an extremely noisy environment as discussed in [Section 3.5](#). A robot with a moderate speed needs to capture more images in an extremely low-light environment than a robot at a high speed. By capturing a bias frame – an expected noise output for each pixel with no light and subtracting it in real-time from each captured image, it is possible to associate the variance of each pixel of the image to the associated thermal noise. A real-time speed control algorithm that has the ability to predict the amount of noise on merged images and by doing so, calculate the motion feasible between frames will improve alignment issues associated with [Chapter 3](#) and the search space described in [Chapter 4](#). We partially address this, by evaluating performance of our methods over different sizes of robotic bursts. While we capture bursts of images from the stream of frames defined by the frame rate of a camera, this can also be controlled actively.

Adaptive threshold. All our methods rely on a key global variable – peak threshold to detect features throughout the images, and across all images for reconstruction. Because captured images have varying contrast across them and illumination could vary between bursts of images for reconstruction, an adaptive threshold is an open research avenue. This can be achieved by either having an optimized lookup table with a sweeping peak threshold for a product of noise and contrast or a weight score variable based on contrast or by using an adaptive loss function.

We anticipate additional advantages from dynamic adjustments of analog camera gain and exposure. However, this raises challenges in terms of motion between frames and partial saturation. This could also be expensive for smaller platforms as they have short flight time and power/weight limitations.

6.2.2 Raw Robotic Burst Imaging

Our focus in this thesis was on developing a new input scheme for correspondence search that can accommodate monochrome or debayered color images. However, using raw images can improve image quality by reducing noise, as highlighted in [\[149\]](#).

Averaging the raw image channels could improve the spatial results of temporally merged images, building on the findings in [Chapter 3](#). Image content from different color channels can also assist in feature extraction of the robotic burst feature finder, especially in low-contrast regions.

Recent learning methods have demonstrated the effectiveness of learning architecture for raw images [[150](#), [151](#)]. These methods could be applied to our final contribution in [Chapter 5](#) to improve denoising and reconstruction in extremely low-light conditions. Furthermore, using raw images provides better control over image-processing parameters, such as white balance and spatial noise reduction, which can lead to an overall improvement in image quality and feature extraction for light-constrained scenes.

6.2.3 Sensor Fusion

We use conventional cameras as our prominent sensor over other sensing modalities for light-constrained reconstruction. While this is efficient and brings more advantages in terms of power and cost for small robotic platforms such as a drone, employing multi-sensor modalities for deployment brings complementary information for robots that have enough power budget.

This opens a range of interesting research problems related to sensor fusion among different camera domains such as fusion between a conventional camera and a RGB-D sensor, thermal and/or a night vision camera, and other modalities like sonar, radar, and [IMU](#). For example, with a calibrated [IMU](#), we can extract the camera position with respect to the other camera poses in a burst which can increase robustness in alignment for [Chapter 3](#) and allow improved slope settings for [Chapter 4](#).

Input data from complementary sensors can also be used as a weak supervisory signal to accelerate the training in extremely low-light conditions and can be leveraged during testing to improve performance.

6.2.4 Broader Impact

Robotic burst imaging for light-constrained 3D reconstruction addresses the current problems in field robotics, but we believe these ideas and developed methods have broader applications beyond robotics.

Burst feature finders demonstrated in [Chapter 4](#) for example can be used for reconstruction of old archaeological sites and caves given the images are captured using similar platform inertia to robots. Similarly, the learning-based approaches can be extended for augmented reality, virtual reality, and other 3D research domains which share keypoint-based reconstruction, i.e., object classification and human pose tracking.

Considering the methods in which we handle motion within bursts of images across all our methods – tile based alignment in [Chapter 3](#), leveraging platform dynamics for 1D and 2D apparent motion in [Chapter 4](#) and generalizing motion in [Chapter 5](#), there are emerging opportunities in the domain of fluid dynamics and microbiology to enable tracking over small motion variation and reconstructing deformable scenes e.g., bursting balloon.

The developed methods enable accurate representation of low-light scenes for reconstruction which may be leveraged for both positive and negative ends. This approach has the potential for privacy misuse, similar to other image-based methods. Recent studies [[152](#), [153](#)] are specifically focusing on addressing privacy concerns associated with image-based 3D reconstruction.

Bibliography

- [1] A. Ravendran, M. Bryson, and D. G. Dansereau, “Burst imaging for light-constrained structure-from-motion,” *Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 1040–1047, 2021.
- [2] —, “BuFF: Burst feature finder for light-constrained 3D reconstruction,” *arXiv preprint arXiv:2209.09470*, 2022.
- [3] Z. Teed and J. Deng, “Droid-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 16 558–16 569, 2021.
- [4] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, “Nice-SLAM: Neural implicit scalable encoding for SLAM,” in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 786–12 796.
- [5] Z. Fu, A. Kumar, A. Agarwal, H. Qi, J. Malik, and D. Pathak, “Coupling vision and proprioception for navigation of legged robots,” in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 273–17 283.
- [6] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, “Vision-only robot navigation in a neural radiance world,” *Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 4606–4613, 2022.
- [7] C. M. Parameshwara, G. Hari, C. Fermüller, N. J. Sanket, and Y. Aloimonos, “Diff-PoseNet: Direct differentiable camera pose estimation,” in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6845–6854.
- [8] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, “Self-supervised 6D object pose estimation for robot manipulation,” in *International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3665–3671.
- [9] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, “Towards real-time monocular depth estimation for robotics: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 16 940–16 961, 2022.
- [10] T.-W. Hui, “RM-Depth: Unsupervised learning of recurrent monocular depth in dynamic scenes,” in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1675–1684.

- [11] M. Johnson-Roberson, M. Bryson, A. Friedman, O. Pizarro, G. Troni, P. Ozog, and J. C. Henderson, “High-resolution underwater robotic vision-based mapping and three-dimensional reconstruction for archaeology,” *Journal of Field Robotics*, vol. 34, no. 4, pp. 625–643, 2017.
- [12] M. Bryson, A. Reid, F. Ramos, and S. Sukkarieh, “Airborne vision-based mapping and classification of large farmland environments,” *Journal of Field Robotics*, vol. 27, no. 5, pp. 632–655, 2010.
- [13] J. Shields, O. Pizarro, and S. B. Williams, “Towards adaptive benthic habitat mapping,” in *International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9263–9270.
- [14] T. Dang, F. Mascarich, S. Khattak, H. Nguyen, H. Nguyen, S. Hirsh, R. Reinhart, C. Papachristos, and K. Alexis, “Autonomous search for underground mine rescue using aerial robots,” in *IEEE Aerospace Conference*, 2020, pp. 1–8.
- [15] B. M. Moghaddam and R. Chhabra, “On the guidance, navigation and control of in-orbit space robotic missions: A survey and prospective vision,” *Acta Astronautica*, vol. 184, pp. 70–100, 2021.
- [16] Y. Wu, X. Ta, R. Xiao, Y. Wei, D. An, and D. Li, “Survey of underwater robot positioning navigation,” *Applied Ocean Research*, vol. 90, p. 101845, 2019.
- [17] J. Maki, D. Gruel, C. McKinney, M. Ravine, M. Morales, D. Lee, R. Willson, D. Copley-Woods, M. Valvo, T. Goodsall, *et al.*, “The Mars 2020 engineering cameras and microphone on the Perseverance rover: A next-generation imaging system for Mars exploration,” *Space Science Reviews*, vol. 216, pp. 1–48, 2020.
- [18] M. Ramezani, K. Khosoussi, G. Catt, P. Moghadam, J. Williams, P. Borges, F. Pauling, and N. Kottege, “Wildcat: Online continuous-time 3D LiDAR-Inertial SLAM,” *arXiv preprint arXiv:2205.12595*, 2022.
- [19] K. Olsson-Francis, P. T. Doran, V. Ilyin, F. Raulin, P. Rettberg, G. Kminek, M. P. Z. Mier, A. Coustenis, N. Hedman, O. Al Shehhi, *et al.*, “The COSPAR Planetary Protection Policy for robotic missions to Mars: A review of current scientific knowledge and future perspectives,” *Life Sciences in Space Research*, 2022.
- [20] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.
- [21] X. Han, H. Laga, and M. Bennamoun, “Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era,” *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 5, pp. 1578–1604, 2021.
- [22] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

- [23] C.-C. Wang, S.-S. Huang, and L.-C. Fu, “Driver assistance system for lane detection and vehicle recognition with night vision,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2005, pp. 3530–3535.
- [24] B. Bird, M. Nancekieveill, A. West, J. Hayman, C. Ballard, W. Jones, S. Ross, T. Wild, T. Scott, and B. Lennox, “Vega — A small, low cost, ground robot for nuclear de-commissioning,” *Journal of Field Robotics*, vol. 39, no. 3, pp. 232–245, 2022.
- [25] A. Krull, T.-O. Buchholz, and F. Jug, “Noise2Void - Learning denoising from single noisy images,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2129–2137.
- [26] J. Batson and L. Royer, “Noise2Self: Blind denoising by self-supervision,” in *International Conference on Machine Learning (ICML)*, 2019, pp. 524–533.
- [27] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
- [28] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, “Deep learning for single image super-resolution: A brief review,” *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.
- [29] A. Levin, P. Sand, T. S. Cho, F. Durand, and W. T. Freeman, “Motion-invariant photography,” *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, pp. 1–9, 2008.
- [30] H. Nagahara, S. Kuthirummal, C. Zhou, and S. K. Nayar, “Flexible depth of field photography,” in *European Conference on Computer Vision (ECCV)*, 2008, pp. 60–73.
- [31] R. Raskar, A. Agrawal, and J. Tumblin, “Coded exposure photography: Motion deblurring using fluttered shutter,” in *ACM SIGGRAPH 2006 Papers*, 2006, pp. 795–804.
- [32] Z. Liu, L. Yuan, X. Tang, M. Uyttendaele, and J. Sun, “Fast burst images denoising,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–9, 2014.
- [33] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, “Burst photography for high dynamic range and low-light imaging on mobile cameras,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.
- [34] O. Liba, K. Murthy, Y.-T. Tsai, T. Brooks, T. Xue, N. Karnad, Q. He, J. T. Barron, D. Sharlet, R. Geiss, *et al.*, “Handheld mobile photography in very low light,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.
- [35] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, “Handheld multi-frame super-resolution,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–18, 2019.

- [36] C. Godard, K. Matzen, and M. Uyttendaele, “Deep burst denoising,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 538–554.
- [37] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, “Deep burst super-resolution,” in *Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9209–9218.
- [38] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” in *European Conference on Computer Vision (ECCV)*, 2014, pp. 834–849.
- [39] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: A versatile and accurate monocular SLAM system,” *IEEE Transactions on Robotics (T-RO)*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [40] H. Bay, T. Tuytelaars, and L. V. Gool, “SURF: Speeded up robust features,” in *European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417.
- [41] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.
- [42] D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperPoint: Self-supervised interest point detection and description,” in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 224–236.
- [43] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, “D2-Net: A trainable CNN for joint description and detection of local features,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8092–8101.
- [44] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, “R2D2: Reliable and repeatable detector and descriptor,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [45] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [46] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, “Stomach 3D reconstruction using virtual chromoendoscopic images,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021.
- [47] X. Chen and M. Sra, “IntoTheVideos: Exploration of dynamic 3D space reconstruction from single sports videos,” in *ACM Symposium on User Interface Software and Technology*, 2021, pp. 14–16.
- [48] H. Alismail, M. Kaess, B. Browning, and S. Lucey, “Direct visual odometry in low light using binary descriptors,” *Robotics and Automation Letters (RA-L)*, vol. 2, no. 2, pp. 444–451, 2016.
- [49] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual odometry,” in *International Conference on Robotics and Automation (ICRA)*, 2014, pp. 15–22.

- [50] C. Kerl, J. Sturm, and D. Cremers, “Dense visual SLAM for RGB-D cameras,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 2100–2106.
- [51] S. Klose, P. Heise, and A. Knoll, “Efficient compositional approaches for real-time robust direct visual odometry from RGB-D data,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 1100–1106.
- [52] F. Steinbrücker, J. Sturm, and D. Cremers, “Real-time visual odometry from dense RGB-D images,” in *International Conference on Computer Vision (ICCV) Workshops*, 2011, pp. 719–722.
- [53] A. Savinykh, M. Kurenkov, E. Kruzhkov, E. Yudin, A. Potapov, P. Karpyshev, and D. Tsetserukou, “DarkSLAM: GAN-assisted visual SLAM for reliable operation in low-light conditions,” in *Vehicular Technology Conference*, 2022, pp. 1–6.
- [54] S. Park, T. Schöps, and M. Pollefeys, “Illumination change robustness in direct visual SLAM,” in *International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4523–4530.
- [55] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperGlue: Learning feature matching with graph neural networks,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947.
- [56] Y.-S. Shin and A. Kim, “Sparse depth enhanced direct thermal-infrared SLAM beyond the visible spectrum,” *Robotics and Automation Letters (RA-L)*, vol. 4, no. 3, pp. 2918–2925, 2019.
- [57] T. Hua, L. Pei, T. Li, Q. Wu, R. Wang, and W. Yu, “I2-SLAM: Fusing infrared camera and IMU for simultaneous localization and mapping,” in *International Conference on Autonomous Unmanned Systems*, 2021, pp. 2834–2844.
- [58] Z. Hong, Y. Petillot, and S. Wang, “RadarSLAM: Radar based large-scale SLAM in all weathers,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5164–5170.
- [59] E. Prassler, A. Ritter, C. Schaeffer, and P. Fiorini, “A short history of cleaning robots,” *Autonomous Robots*, vol. 9, pp. 211–226, 2000.
- [60] A. K. Pandey and R. Gelin, “A mass-produced sociable humanoid robot: Pepper: The first machine of its kind,” *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.
- [61] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, “Zero-reference deep curve estimation for low-light image enhancement,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1780–1789.
- [62] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, “From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3063–3072.

- [63] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5637–5646.
- [64] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2016.
- [65] D. Jin, L. Ma, R. Liu, and X. Fan, "Bridging the gap between low-light scenes: Bilevel learning for fast adaptation," in *ACM International Conference on Multimedia*, 2021, pp. 2401–2409.
- [66] A. Lengyel, S. Garg, M. Milford, and J. C. Van Gemert, "Zero-shot day-night domain adaptation with a physics prior," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 4399–4409.
- [67] X. Deng, P. Wang, X. Lian, and S. Newsam, "Nightlab: A dual-level architecture with hardness detection for segmentation at night," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 938–16 948.
- [68] Z. Zheng, Y. Wu, X. Han, and J. Shi, "ForkGAN: Seeing into the rainy night," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 155–170.
- [69] A. Punnappurath, A. Abuolaim, A. Abdelhamed, A. Levinshtein, and M. S. Brown, "Day-to-night image synthesis for training nighttime neural ISPs," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 769–10 778.
- [70] K. Monakhova, S. R. Richter, L. Waller, and V. Koltun, "Dancing under the stars: Video denoising in starlight," in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 241–16 251.
- [71] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A* 3D dataset: Towards autonomous driving in challenging environments," in *International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2267–2273.
- [72] T. Gruber, M. Bijelic, F. Heide, W. Ritter, and K. Dietmayer, "Pixel-accurate depth evaluation in realistic driving scenarios," in *International Conference on 3D Vision (3DV)*, 2019, pp. 95–105.
- [73] R. Szeliski, *Computer vision: Algorithms and applications*. Springer Science & Business Media, 2010.
- [74] Z. Zhao, Y. Zhai, B. M. Chen, and P. Liu, "BALF: Simple and efficient blur aware local feature detector," *arXiv preprint arXiv:2211.14731*, 2022.
- [75] P. Liu, X. Zuo, V. Larsson, and M. Pollefeys, "MBA-VO: Motion blur aware visual odometry," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 5550–5559.

- [76] D. Jayaraman and K. Grauman, “Slow and steady feature analysis: Higher order temporal coherence in video,” in *Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3852–3861.
- [77] H.-J. Chien, C.-C. Chuang, C.-Y. Chen, and R. Klette, “When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry,” in *International Conference on Image and Vision Computing New Zealand*, 2016, pp. 1–6.
- [78] G. Csurka and M. Humenberger, “From handcrafted to deep local invariant features,” *arXiv preprint arXiv:1807.10254*, vol. 2, p. 1, 2018.
- [79] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [80] I. Stanković, M. Brajović, J. Lerga, M. Daković, and L. Stanković, “Image denoising using RANSAC and compressive sensing,” *Multimedia Tools and Applications*, pp. 1–23, 2022.
- [81] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [82] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment - A modern synthesis,” in *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms*, 2000, pp. 298–372.
- [83] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000 km: The Oxford RobotCar dataset,” *International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [84] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, *et al.*, “Benchmarking 6DOF outdoor visual localization in changing conditions,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8601–8610.
- [85] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, and P. Newman, “Real-time kinematic ground truth for the Oxford RobotCar dataset,” *arXiv preprint arXiv:2002.10152*, 2020.
- [86] M. W. Tao, J. Malik, and R. Ramamoorthi, “Sharpening out of focus images using high-frequency transfer,” in *Computer Graphics Forum*, vol. 32, no. 2pt4. Wiley Online Library, 2013, pp. 489–498.
- [87] A. Meras, “Two passions: Tennis and photography,” *PSA Journal*, vol. 79, no. 12, pp. 34–39, 2013.
- [88] S. W. Hasinoff, “Photon, Poisson noise.” *Computer Vision, A Reference Guide*, vol. 4, 2014.

- [89] G. E. Healey and R. Kondepudy, “Radiometric CCD camera calibration and noise estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 3, pp. 267–276, 1994.
- [90] A. C. Bovik, *The essential guide to image processing*. Academic Press, 2009.
- [91] E. R. Davies, *Computer and machine vision: Theory, algorithms, practicalities*. Academic Press, 2012.
- [92] R. G. Driggers, R. Vollmerhausen, and K. Krapels, “Target identification performance as a function of temporal and fixed pattern noise,” *Optical Engineering*, vol. 40, no. 3, pp. 443–447, 2001.
- [93] S. C. Cain, M. M. Hayat, and E. E. Armstrong, “Projection-based image registration in the presence of fixed-pattern noise,” *IEEE Transactions on Image Processing*, vol. 10, no. 12, pp. 1860–1872, 2001.
- [94] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, “Noise2Noise: Learning image restoration without clean data,” in *International Conference on Machine Learning (ICML)*, 2018, pp. 2965–2974.
- [95] Y. Quan, M. Chen, T. Pang, and H. Ji, “Self2Self with dropout: Learning self-supervised denoising from single image,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1890–1898.
- [96] M. Tassano, J. Delon, and T. Veit, “FastDVDnet: Towards real-time deep video denoising without flow estimation,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1354–1363.
- [97] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, “Supervised raw video denoising with a benchmark dataset on dynamic scenes,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2301–2310.
- [98] Y. Nan, Y. Quan, and H. Ji, “Variational-EM-based deep learning for noise-blind image deblurring,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3623–3632.
- [99] J. Pan, H. Bai, and J. Tang, “Cascaded deep video deblurring using temporal sharpness prior,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3043–3051.
- [100] Y. Yuan, W. Su, and D. Ma, “Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3555–3564.
- [101] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, “Multi-scale boosted dehazing network with dense feature fusion,” in *International Conference on Computer Vision (ICCV)*, 2020, pp. 2157–2167.

- [102] Y. Li, Q. Miao, W. Ouyang, Z. Ma, H. Fang, C. Dong, and Y. Quan, “LAP-Net: Level-aware progressive network for image dehazing,” in *International Conference on Computer Vision (ICCV)*, 2019, pp. 3276–3285.
- [103] X. Liu, Y. Ma, Z. Shi, and J. Chen, “GridDehazeNet: Attention-based multi-scale network for image dehazing,” in *International Conference on Computer Vision (ICCV)*, 2019, pp. 7314–7323.
- [104] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang, “Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6306–6314.
- [105] V. Sharma, A. Diba, D. Neven, M. S. Brown, L. Van Gool, and R. Stiefelhagen, “Classification-driven dynamic image enhancement,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4033–4041.
- [106] R. Wang, Q. Zhang, C. W. Fu, X. Shen, W. S. Zheng, and J. Jia, “Underexposed photo enhancement using deep illumination estimation,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6849–6857.
- [107] K. Wei, Y. Fu, J. Yang, and H. Huang, “A physics-based noise formation model for extreme low-light raw denoising,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2758–2767.
- [108] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, “How to train your robot with deep reinforcement learning; Lessons we’ve learned,” *International Journal of Robotics Research (IJRR)*, 2021.
- [109] H. S. Choi, C. Crump, C. Duriez, A. Elmquist, G. Hager, D. Han, F. Hearl, J. Hodgins, A. Jain, F. Leve, C. Li, F. Meier, D. Negrut, L. Righetti, A. Rodriguez, J. Tan, and J. Trinkle, “On the use of simulation in robotics: Opportunities, challenges, and suggestions formoving forward,” *Proc. of the National Academy of Sciences of the United States of America*, vol. 118, no. 1, pp. 1–9, 2021.
- [110] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [111] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” *Electronics*, vol. 8, no. 8, p. 832, 2019.
- [112] S. W. Hasinoff, F. Durand, and W. T. Freeman, “Noise-optimal capture for high dynamic range photography,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 553–560.
- [113] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, “Burst denoising with kernel prediction networks,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2502–2510.

- [114] S. Ma, S. Gupta, A. C. Ulku, C. Bruschini, E. Charbon, and M. Gupta, “Quanta burst photography,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 79–1, 2020.
- [115] S. Zhu, R. Zhang, L. Zhou, T. Shen, T. Fang, P. Tan, and L. Quan, “Very large-scale global SfM by distributed motion averaging,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4568–4577.
- [116] M. Delbracio and G. Sapiro, “Burst deblurring: Removing camera shake through Fourier burst accumulation,” in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2385–2393.
- [117] D. G. Dansereau, B. Girod, and G. Wetzstein, “LiFF: Light field features in scale and depth,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8042–8051.
- [118] M. Rad, M. Oberweger, and V. Lepetit, “Feature mapping for learning fast and accurate 3D pose inference from synthetic images,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4663–4672.
- [119] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.
- [120] J. R. Janesick, *Scientific charge-coupled devices*. SPIE press, 2001, vol. 83.
- [121] J. Hu, O. Gallo, K. Pulli, and X. Sun, “HDR deghosting: How to deal with saturation?” in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1163–1170.
- [122] M. Bansal, M. Kumar, and M. Kumar, “2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors,” *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18 839–18 857, 2021.
- [123] P. Scovanner, S. Ali, and M. Shah, “A 3-Dimensional SIFT descriptor and its application to action recognition,” in *ACM International Conference on Multimedia*, 2007, pp. 357–360.
- [124] M. Al Ghamdi, L. Zhang, and Y. Gotoh, “Spatio-temporal SIFT and its application to human action classification,” in *European Conference on Computer Vision (ECCV)*, 2012, pp. 301–310.
- [125] S. Yu, B. Park, J. Park, and J. Jeong, “Joint learning of blind video denoising and optical flow estimation,” in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.
- [126] D. G. Dansereau, O. Pizarro, and S. B. Williams, “Linear volumetric focus for light field cameras,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, pp. 15–1, 2015.
- [127] K. He, Y. Lu, and S. Sclaroff, “Local descriptors optimized for average precision,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 596–605.

- [128] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, “HPatches: A benchmark and evaluation of handcrafted and learned local descriptors,” in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5173–5182.
- [129] P. Di Febbo, C. Dal Mutto, K. Tieu, and S. Mattoccia, “KCNN: Extremely-efficient hardware keypoint detection with a compact convolutional neural network,” in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 682–690.
- [130] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision (ECCV)*, 2006, pp. 430–443.
- [131] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, “Key. Net: Keypoint detection by handcrafted and learned CNN filters,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5836–5844.
- [132] W. Hartmann, M. Havlena, and K. Schindler, “Predicting matchability,” in *Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 9–16.
- [133] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [134] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [135] Z. Huo, Y. Zhang, H. Liu, J. Wang, X. Liu, and J. Zhang, “Improved covariant local feature detector,” *Pattern Recognition Letters*, vol. 135, pp. 1–7, 2020.
- [136] X. Zhang, F. X. Yu, S. Karaman, and S.-F. Chang, “Learning discriminative and transformation covariant local feature detectors,” in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6818–6826.
- [137] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, “Comparative evaluation of hand-crafted and learned local features,” in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1482–1491.
- [138] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, “MatchNet: Unifying feature and metric learning for patch-based matching,” in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3279–3286.
- [139] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, “Discriminative learning of deep convolutional feature point descriptors,” in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 118–126.
- [140] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, “Universal correspondence network,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.
- [141] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, “PN-Net: Conjoined triple deep network for learning local image descriptors,” *arXiv preprint arXiv:1601.05030*, 2016.

- [142] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, “Working hard to know your neighbor’s margins: Local descriptor learning loss,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [143] X. Zhang, F. X. Yu, S. Kumar, and S.-F. Chang, “Learning spread-out local feature descriptors,” in *International Conference on Computer Vision (ICCV)*, 2017, pp. 4595–4603.
- [144] Y. Tian, B. Fan, and F. Wu, “L2-Net: Deep learning of discriminative patch descriptor in Euclidean space,” in *Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 661–669.
- [145] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “LIFT: Learned invariant feature transform,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 467–483.
- [146] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, “Large-scale image retrieval with attentive deep local features,” in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3456–3465.
- [147] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Revisiting Oxford and Paris: Large-scale image retrieval benchmarking,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5706–5715.
- [148] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [149] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, “Unprocessing images for learned raw denoising,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 036–11 045.
- [150] T. Ehret, A. Davy, P. Arias, and G. Facciolo, “Joint demosaicking and denoising by fine-tuning of bursts of raw images,” in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8868–8877.
- [151] L. Liu, X. Jia, J. Liu, and Q. Tian, “Joint demosaicing and denoising with self guidance,” in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2240–2249.
- [152] C. Hinojosa, J. C. Niebles, and H. Arguello, “Learning privacy-preserving optics for human pose estimation,” in *International Conference on Computer Vision (ICCV)*, 2021, pp. 2573–2582.
- [153] M. Geppert, V. Larsson, J. L. Schönberger, and M. Pollefeys, “Privacy preserving partial localization,” in *Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 337–17 347.