Advancing Forest Health Monitoring: Harnessing the Power of Deep Learning Computer Vision for Remote Sensing Applications

by

Rudraksh Kapil

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Forests provide immense economic, ecological, and societal values, making forest health monitoring (FHM) a crucial task for guiding conservation and management of these essential ecosystems. Drones have seen increased popularity in this domain due to their ability to collect high-resolution, multi-modal images over a large area of interest (AOI). Naturally, different sensors (e.g., thermal) can capture more information than just RGB cameras and lead to a more comprehensive understanding of the AOI. The processing and analysis of these images has largely been done manually or using manually crafted indices, posing a severe bottleneck in terms of the size of the AOI and generalizability of results to different locations with dissimilar tree species. Computer vision techniques, particularly those relying on deep learning (DL), have the potential to overcome these issues and yield more effective FHM, especially when information from multiple sensors is combined. Therefore, the overarching goal of this thesis is successfully applying DL and computer vision techniques to process and analyze multi-modal drone images for FHM.

Towards achieving this goal, first, a new workflow to generate high-quality thermal orthomosaics is proposed. Orthomosaicking removes distortions from nadir (i.e., downward-facing) images and stitches them together to produce one broader image encompassing the entire AOI. Typical thermal-only orthomosaicking workflows suffer from gaps and swirling artifacts due to the poor structure-from-motion (SfM) performance on the low-contrast and low-resolution thermal images. Instead, the proposed workflow leverages the superior SfM results from simultaneously acquired, higher-quality RGB images and performs image co-registration using a learned affine

transformation to generate thermal orthomosaics that are free from the mentioned issues and precisely aligned with their RGB counterparts, without disturbing the radiometric information of the original images. Second, the focus shifts to precisely detecting individual tree crowns from the aligned RGB-thermal imagery. Shorter trees hidden in RGB images by the shadows of neighbouring larger trees become apparent in thermal images. Detecting these trees correctly is critical in many monitoring tasks, e.g., bark beetles preferentially attack smaller, younger trees during their endemic population stages. To appropriately leverage both image modalities, a novel unsupervised domain adaptation (UDA) strategy is proposed to adapt an existing state-of-the-art RGB-only detection model to thermal data and fuse the features extracted from both prior to detection. The proposed method outperforms existing UDA and image-level fusion techniques without requiring any annotations for training. Finally, the vital FHM task of bark beetle attack stage classification is considered. In sufficiently large numbers, these insects pose a devastating threat to forest ecosystems by exacerbating tree mortality. Infested trees gradually show crown discoloration in four separate 'attack' stages, and effectively distinguishing between these stages over a wide area can drastically expedite the early detection of bark beetle outbreaks. Traditionally, manual identification is done by experts using helicopter surveys or collected imagery, both of which are arduous tasks. Instead, the proposed method in this thesis leverages a transfer learning technique to train a deep attack-stage classification model that distinguishes between all visible stages with a near-perfect accuracy in the presence of limited training data.

Across all three objectives, the novel methods proposed in this thesis show significant improvement over previous state-of-the-art techniques. These results are derived through extensive experimentation on different datasets. For the first two objectives, a newly collected RGB-thermal drone image dataset over a forested region in central Alberta, Canada, is used. For the third, an existing bark beetle attack stage classification dataset collected from a forested region in Northern Mexico is used.

Preface

The central chapters of this thesis are based on papers that are either published or currently under review. Chapters 2 and 4 are each written based on separate papers that have previously been published [1, 3], while Chapter 3 is based on a paper that is currently under review [2].

[1] **Chapter 2**: R. Kapil, G. Castilla, S. M. Marvasti-Zadeh, D. Goodsman, N. Erbilgin, and N. Ray, "Orthomosaicking thermal drone images of forests via simultaneously acquired RGB images," Feature Paper in MDPI Journal of Remote Sensing, vol. 15, no. 10, p. 2653, (2023). DOI: 10.3390/rs15102653.

[2] **Chapter 3**: R. Kapil, S. M. Marvasti-Zadeh, N. Erbilgin, and N. Ray, "ShadowSense: unsupervised domain adaptation and feature fusion for shadow-agnostic tree crown detection from RGB-thermal drone imagery," (2023). Under review IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

[3] **Chapter 4**: R. Kapil, S. M. Marvasti-Zadeh, D. Goodsman, N. Ray, and N. Erbilgin, "Classification of bark beetle-induced forest tree mortality using deep learning," (2022). DOI: 10.48550/arXiv.2207.07241. Presented at Visual observation and analysis of Vertebrate And Insect Behavior Workshop held at the 26th International Conference on Pattern Recognition (ICPR 2022).

I am the primary contributor of all three papers, having handled the design, implementation, experiments, analysis, and initial manuscript preparation. Profs. Nilanjan Ray and Nadir Erbilgin co-supervised all papers and helped with revisions. Dr. Seyed Mojtaba Marvasti-Zadeh contributed to the review and editing for all three and is listed as an equal primary contributor for the third paper. Dr. Guillermo Castilla contributed to the review and editing of the first paper. In addition, all listed co-authors contributed to the ideation and revisions for each paper.

To my uncle Amit Sharma, in loving memory.

You were an ardent advocate for my higher education,

and I will always appreciate your encouragement.

Acknowledgements

First and foremost, I express my deepest gratitude to both of my co-supervisors, Prof. Nilanjan Ray from the Department of Computing Science and Prof. Nadir Erbilgin from the Department of Renewable Resources (REN R). Their unwavering support, invaluable guidance, and constant belief in my ideas were an integral part of my degree, and I will forever be grateful to them. I immensely thank my other collaborators throughout this project as well – Dr. Seyed Mojtaba Marvasti-Zadeh of REN R, Dr. Guillermo Castilla of Natural Resources Canada (NRCan), and Dr. Devin Goodsman of NRCan. The publications resulting from this project would not have been possible without their expertise and irreplaceable contributions.

This research project was funded by fRI Research-Mountain Pine Beetle Ecology Program. In addition, I thank the Weyerhauser Company[1], holder of the Forest Management Agreement, for kindly granting access to the Cynthia cutblock area and permission for collecting data. I also thank Michael Gartrell, Jim Weber, and Steven Wagers of NRCan for contributing to drone mission planning and data acquisition.

Finally, I am eternally indebted to my amazing family: my parents for their unwavering love, encouragement, and sacrifices throughout my lifelong educational pursuits; my sister Vaishnavi for inspiring me and always guiding me down the right path; and my partner Anam for her care, understanding, and optimism during what has been the most challenging step of my academic journey so far. All of them have been my fundamental pillars of support, being there for me no matter what. Without them, I could not have achieved my goals and reached where I am today.

[1] https://www.weyerhaeuser.com/

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis Statement 5

1.3 Contributions 5

1.4 Thesis Layout 8

2 RGB-Thermal Orthomosaicking Pipeline **10**

2.1 Introduction and Related Work 10

2.2 Materials and Proposed Workflow 16

2.2.1 Cynthia Cutblock Study Site 16

2.2.2 Proposed Integrated Orthomosaicking Workflow 17

2.2.3 Downstream ITCD Task 29

2.2.4 Proposed Open-Source Tool 30

2.2.5 Performance Assessment 31

2.3 Experimental Results 32

2.3.1 Quantitative Results 32

2.3.2 Qualitative Results 35

2.3.3 Robustness of Transformation Matrix Computation 37

2.3.4 Radiometric Analysis 39

2.3.5 Downstream Task Performance 40

2.3.6 Processing Time 41

2.4 Discussion 41

2.5 Conclusions and Future Work 45

3 Shadow-Agnostic Tree Crown Detection **47**

3.1 Introduction 47

3.2 Related Work 50

3.3 Proposed Method and Dataset 52

3.3.1 Model Architecture and Training 53

3.3.2 Feature Fusion during Inference . . . 56
3.3.3 Dataset for Shadowed Tree Crown Detection . . . 58
3.4 Experiments . . . 59
3.4.1 Implementation Details . . . 60
3.4.2 Baseline Quantitative Comparison . . . 61
3.4.3 State-of-the-art Quantitative Comparison . . . 61
3.4.4 Ablation Study . . . 64
3.4.5 Qualitative Results . . . 65
3.5 Conclusions . . . 66

4 Bark Beetle Visible Attack Stage Classification 68
4.1 Introduction . . . 68
4.2 Related Works . . . 70
4.3 Proposed Method . . . 71
4.4 Empirical Evaluations . . . 74
4.4.1 Implementation Details . . . 74
4.4.2 Experimental Results . . . 76
4.4.3 Ablation Study . . . 77
4.5 Conclusion . . . 78

5 Conclusions 79
5.1 Summary of Contributions . . . 79
5.2 Implications . . . 83
5.3 Future Work . . . 84

Bibliography 86

Appendix A: RT-Trees Dataset Additional Information 100

Appendix B: ShadowSense Extended Ablation Study 107

List of Tables

2.1 **Summary of Cynthia Cutblock Data for each Flight Date**. Temperature and humidity values are taken from three weather stations closest to the cutblock and averaged. 18

2.2 **Average MI of Individual RGB and Thermal Images** obtained using different design choices in the proposed workflow for five flights over the Cynthia cutblock. The best values are emboldened. 35

2.3 **Robustness of NGF-based Co-registration**. Average, minimum, and maximum observed values over all flights for each affine transformation M component are listed. The coefficient of variation (CoV) for each component is reported in the final column. $M_{i,j}$ denotes the value in the ith row and jth column. $M_{3,1}$ and $M_{3,1}$ are always 0 for affine transformation matrices and thus omitted. 38

3.1 **Categorization of Related Works** according to training supervision through RGB ground truth (GT) annotations. 52

3.2 **Comparative Overview** of RGB-Thermal Image Datasets. 59

3.3 **Quantitative Comparison** of the proposed method with baseline and SOTA methods using % AP50 (↑), % AR100 (↑), and % of shadowed trees correctly identified (↑). Best and second-best results are emboldened in red (supervised) and blue (self-supervised). 62

4.1 **Dataset Distribution** for each flight according to attack stage label and training/validation/testing split. 75

4.2 **Classification Accuracy** for various attack stage classification models. The best result is emboldened. 77

A.1 **RT-Trees Dataset Information by Flight Date**. All dates are from the year 2022. Information about the flight, lighting and weather conditions, and number of images is listed. Approximately 70% of the raw image pairs captured for a given date are sampled for the training set based on GPS location (see Fig. A.2), and then divided into six 500×500 patches. From the August 30 data, 63 images are taken for testing and 10 for validation, hence the total number of image pairs in RT-Trees is 49879. 101

B.1 **Extended Ablation Study** for different hyperparameter settings in the proposed method based on AP50 and AP100 metrics, trained without annotation on the RT-trees training set. Results on the RT-Trees validation set are reported. While changing one hyperparameter, all others are set to their best-performing values as described in the implementation details for the proposed ShadowSense configuration (also emboldened here). 108

List of Figures

1.1 **Overview of Thesis Objectives** that collectively constitute a complete RS pipeline. Novel contributions are set forth for each objective in this thesis, as presented in the central chapters. 6

2.1 Comparison of **(a) Existing Workflows** for thermal orthomosaic generation and **(b) The Proposed Workflow** that leverages intermediate outputs from RGB orthomosaic generation. Thermal-only processing workflows are prone to gaps and swirling artifacts (shown in red), which are tackled by the proposed workflow. 12

2.2 **Left**: Location of Cynthia cutblock in Alberta, Canada. **Right**: A close-up of the area including the village of Cynthia. 18

2.3 **Overview of the Integrated RGB-thermal Orthomosaicking Workflow**. Example images of each type of data are also included, connected by dashed lines. Stage 3 is more detailed in Fig. 2.4. . . . 19

2.4 **Detailed Steps of the Automated Intensity-based Image Co-registration**. After preprocessing the images, K RGB and thermal image pairs are systematically sampled for batch processing. The parameters of the Homography Module [92] represent the values in the transformation matrix being computed and are learned during gradient descent optimization [106] of the normalized gradient fields (NGF) loss [60] between each transformed thermal Gaussian pyramid and its corresponding RGB pyramid. 24

2.5 **Internal Mechanism of the Homography Module**, which leverages matrix exponential to compute the affine transformation matrix M using 6 basis matrices B and learnable scalar parameters v. The same 3×3 matrix C is used to update A at each of the 10 substeps, which are finally summed to yield M. The term *eye(3)* denotes the identity matrix of size 3×3. 27

2.6 **Graphical User Interface** for the orthomosaicking tool showing the various easily-configurable processing options. 31
2.7 **Mutual Information (MI) for August 30 Cynthia Flight**. Box (interquartile range, IQR) and whisker (within 1.5× IQR) plots are shown for the MI between 814 RGB-thermal image pairs using different co-registration techniques. NGF is normalized gradient fields [34], while ECC is enhanced correlation coefficient [23]. The white circles denote mean values. 34
2.8 **Visualization of Orthomosaics** generated from August 30 Cynthia cutblock data. (a) Georeferenced RGB orthomosaic. (b) Georeferenced thermal orthomosaic from unregistered images. (c) Georeferenced thermal orthomosaic from NGF-registered images. The improved quality of the orthomosaic in (c) is especially evident from the circular inset showing a straight path between the trees (similar to (a)) compared to the jagged path in (b). 36
2.9 **Checkerboard Visualization** of two samples of undistorted RGB images interlaced with their corresponding thermal images (a) before and (b) after performing image co-registration with the proposed workflow. Coloured squares correspond to the RGB images, and grayscale ones to the thermal images. 37
2.10 **Visual and Radiometric Similarity** between thermal images and the thermal orthomosaic. (a) A thermal image, (b) its corresponding patch in the thermal orthomosaic generated from the proposed workflow, and (c) temperature histograms for both in equally-spaced bins of $0.1°C$. 39
2.11 **Visualization of Downstream ITCD Task Performance on Orthomosaics** generated from the August 30 Cynthia data. (a) Detected tree crowns from the RGB orthomosaic. The percentages represent the detection model's confidence score for each RGB crown. (b) Corresponding patches extracted from the thermal orthomosaic at the same pixel locations as in (a). 40

3.1 **Overview of Proposed Method**. Undetected trees hidden in shadows are indicated by dotted red boxes. Best viewed in color. 48

3.2 **(a) Detailed Workflow of Proposed Training Procedure** consisting of a thermal branch (in red) and an RGB branch (in blue). The weights of both are initialized from [136], and the RGB branch is frozen during training. The thermal feature extractor is trained to fool the domain discriminators (in green), and vice versa, using gradient reversal layers (GRL) at multiple levels. **(b) Close-up** of FPN feature alignment (in purple) at the M3 level that encourages foreground feature map regions of the two branches for a given image pair to match. . . 53
3.3 **Masked Fusion During Inference** for the M2 level feature maps as an example. Background features (purple) are obtained by weighted averaging of the RGB (blue) and thermal (red) features. Foreground features are assigned the original RGB values. Best viewed in colour. 57
3.4 **t-SNE Visualization** of RGB-thermal FPN features: (top row) before training and (bottom row) after training. 65
3.5 **Tree Crown Detection Results**. Each column shows (a) RGB image, (b) Thermal image, (c) Generated mask; and predictions by (d) Baseline [136], (e) DAT-adapted thermal branch, (f) Proposed ShadowSense, and (e) Ground truth. Best viewed in colour. 67

4.1 **Typical Life Cycle of Bark Beetles** and their effect on host tree foliage over time. Beetle images have been adapted from [109]. 69
4.2 **Brief Diagram** of the desired classification model for this task. . . . 70
4.3 **RGB Colour Space Distribution of Bark Beetle Dataset Images**. The borders of the highlighted challenging samples indicate their true labels. 72
4.4 **Histograms** showing RGB colour space distribution of the different attack stages with leaves. 72
4.5 **An Overview of the Proposed Method**. First, the ResNet-50 and feature pyramid network (FPN) are initialized using the tree crown detection pre-trained baseline weights [136]. Following that, the network is modified and fine-tuned to classify the stages of bark beetle attack. 73
4.6 **Data Processing Pipeline** for the bark beetle attack stage classification orthomosaics. 75
4.7 **Visualization of Data Augmentation Strategies** considered to produce minority class samples. 75
4.8 **Confusion Matrices** for the best-performing proposed model. . . . 76
4.9 **t-SNE Visualization** of dataset with different augmentations. . . . 78

A.1 **Kernel Density Estimation Plot for Average Brightness** for images in each flight date, mapped first to LAB color space. 102
A.2 **GPS-based Data Split**. Each point represents the location where a drone image was taken. The assigned cutoff line separates the testing area from the training (and validation) area. 103
A.3 **Splitting Training and Evaluation Images** into patches. All training images are evenly split into six patches, whereas every third evaluation image (testing & validation sets) is centre-cropped. 104
A.4 **Distribution of Bounding Boxes per Image** for all boxes (top) and only difficult boxes (bottom) in the testing set. 104
A.5 **Distribution of Bounding Boxes Areas** for all boxes, difficult boxes only, and non-difficult boxes (i.e., visible in RGB image) in the testing set. 105
A.6 **Distribution of Bounding Boxes Dimensions** for all boxes in the testing set. 105
A.7 **Example of Drone-collected Image Pairs** for each flight date after performing co-registration. 106

Abbreviations

AOI Area of Interest.

BG Background.

CNN Convolutional Neural Network.

CoV Coefficient of Variation.

DAT Domain Adversarial Training.

DL Deep Learning.

DNN Deep Neural Network.

DSM Digital Surface Model.

ECC Enhanced Correlation Coefficient.

EXIF Exchangeable Image File Format (header).

FG Foreground.

FHM Forest Health Monitoring.

FOV Field of View.

FPN Feature Pyramid Network.

GIS Geographic Information System.

GRL Gradient Reversal Layer.

GSD Ground Sampling Distance.

GUI Graphical User Interface.

ITCD Individual Tree Crown Detection.

JPEG Joint Photographic Experts Group (image format).

KNN K Nearest Neighbors.

MI Mutual Information.

ML Machine Learning.

MPB Mountain Pine Beetle.

NGF Normalized Gradient Fields.

ODM Open Drone Map.

RF Random Forest.

RJPEG Radiometric JPEG.

RS Remote Sensing.

SfM Structure from Motion.

SOTA State-of-the-art.

SVM Support Vector Machine.

TIFF Tag Image File Format.

UDA Unsupervised Domain Adaptation.

Glossary of Terms

Geographic Information System A program that can be used to display and analyze georeferenced information like orthomosaics, using data corresponding to a distinct location.

Georeferencing When the internal coordinate system of an orthomosaic can be related to a ground system of geographic coordinates.

Image Registration The process of aligning and overlaying two or more images of the same scene or object by estimating a transformation matrix that captures the spatial relationship between the images.

Nadir Image An aerial photo of the Earth taken vertically downards (i.e., a drone flying above an object taking pictures of it directly from above).

Orthomosaic Photogrammetrically ortho-rectified image product mosaicked from an image collection, where the geometric distortion has been corrected and the imagery has been color balanced to produce a seamless mosaic dataset.

Orthorectification The process of removing image distortions or displacements caused by sensor tilt and topographic relief. The aim is to ensure that every point on the image is represented as if it were captured directly below the sensor (i.e., at nadir).

Photogrammetry The technique of extracting 3D measurements and geometric information from 2D images.

Transfer Learning The process of utilizing knowledge gained from one task or domain to improve the performance on a different but related task or domain.

Unsupervised Domain Adaptation The process of adapting a machine learning model trained on a labeled source domain to perform well on an unlabeled target domain.

Chapter 1

Introduction

1.1 Motivation

Drones can efficiently capture close-range, high-resolution imagery that can provide an aerial view over a large area of interest (AOI). Given recent advancements in remote sensing (RS) technologies, drone imagery has seen extensive utilization within various application domains, including forestry RS, precision agriculture, and infrastructure inspection [150]. Forestry RS, in particular, comprises numerous sub-areas [30], such as estimating forest structural parameters like tree height and canopy area [24], tree species classification [151], forest fire monitoring [91], and deforestation assessment [98]. Among the sub-areas of forestry RS, forest health monitoring (FHM) is crucial for conserving and managing these essential ecosystems [104]. Diseases and pest infestations must be identified early to mitigate their spread through forests [83]. In most FHM tasks, forest managers require an overhead view of AOIs for conducting effective, large-scale analysis and often utilize drones for this purpose [21].

An alternative to drone-based RS for common FHM applications involves using satellites. However, this is subpar for FHM at the individual tree level [30]. The substantial altitude difference between orbiting satellites and AOIs on Earth's surface yields a lower spatial resolution in satellite imagery, making identifying individual trees more difficult than when using drones. Moreover, satellites are more expensive yet less flexible in operation and manoeuvrability due to their slow temporal resolutions [95].

Modern drone systems, on the other hand, can be programmed with specific flight paths according to the need of practitioners, thereby allowing for more flexibility in applications [21]. Granted favourable weather conditions, drones can be easily flown over AOIs whenever required, and their closer proximity to the AOI yields images of higher resolution than with satellites [150]. Predetermined flight paths are especially useful for standardized repeated acquisitions for the same AOI to track changes over time [102]. Because of these advantages, drone-based RS was chosen for the studies conducted as a part of this thesis.

By leveraging synchronized, multiple-sensor cameras mounted aboard drones, a more comprehensive representation of an AOI can be acquired [21]. Commonly used RGB sensors capture visible spectrum information, i.e., the part of the spectrum that humans can also perceive, while thermal sensors are more versatile and can capture near-infrared radiation (NIR) as well as short-, medium-, and long-wave infrared radiation (IR) that is otherwise invisible to humans. Drones can carry various other types of sensors, such as multispectral that capture several separated bands of wavelengths (e.g., red, green, blue, NIR, panchromatic, etc.) or hyperspectral that provide spectral information from thousands of sequential, narrow bands [1]. Working with disparate data modalities poses significant challenges – the difference in sensor specifications yields different image dimensions, resolutions, and fields of view that can cause the same object to appear dissimilar between images captured by different sensors. Despite these challenges, the complementary spectral information from multiple sensors has benefited several drone-based FHM applications [30]. RGB-thermal imagery has been particularly useful for forest fire monitoring [71] and detecting insect-induced canopy temperature increases [118]. Still, there remains ample room for improvement over existing techniques for preprocessing and analyzing multi-modal drone-collected images. In this thesis, I have chosen to highlight this gap in the case of RGB-thermal imagery and demonstrate the benefit of utilizing both of these modalities for the RS tasks of orthomosaicking and individual tree crown detection (ITCD).

During a single drone flight over an AOI, several hundreds or even thousands of images are captured with a high degree of overlap between successive images ($\sim 80\%$). To reduce the redundancy in these images and provide a more comprehensive representation of the area of interest, the images are typically 'stitched' together into a larger image than can be viewed in Geographic Information Systems (GIS). This is called orthomosaicking and is a common preprocessing step in many drone-based RS applications. Orthomosaicking RGB images can be effectively accomplished using existing techniques, such as Open Drone Map [97], since these images typically have the high resolution and contrast that is required for the photogrammetry algorithms performed during orthomosaicking. However, orthomosaicking thermal drone images can pose a significant challenge, especially in forest environments [19, 44, 79, 105]. Thermal images have lower resolution and contrast, making them less suitable for standard orthomosaicking workflows. Thus, existing workflows are typically only able to generate low-quality orthomosaics with incomplete AOI coverage and significant local distortions [138]. Because of these issues, previous FHM works either limit themselves to the narrow visible slice of the EM spectrum captured in RGB imagery (e.g., [96]) or forego orthomosaic generation when working with multi-modal imagery, choosing instead to work with individual images directly (e.g., [80, 118]). The latter can lead to complications when transferring the information to GIS (i.e., ArcGIS or QGIS), which excel at visualizing spatial data with precise global positioning and are thus heavily used for FHM applications [4].

Once drone imagery has been acquired and appropriately preprocessed (i.e., orthomosaicked), further analysis has been typically done either through laborious manual inspection or using manually-defined, rule-based analyses involving simple operations (e.g., cellular automaton [113] and various spectral vegetation indices [6, 122]). The former is prone to human subjectivity, while the latter may not generalize on a larger scale and often needs to be devised using expert domain knowledge. Thus, there is a clear need for more robust autonomous systems to mitigate such subjectivity,

improve reliability, and facilitate deployment over larger AOIs. Existing works that use automated computer vision and image processing analysis have often limited themselves to classical machine learning (ML) algorithms (e.g., semantic segmentation with support vector machines in [9] or a variant of k-nearest neighbours in [51]) that are ill-suited to raw images and often require hand-crafted feature extraction using prior domain knowledge. Deep learning (DL) methods bypass this problem by allowing models to learn how to extract task-relevant feature representations by themselves during training. DL has seen success in RS applications such as tree crown detection [136] and pest detection [88, 93, 108]. Still, as I will demonstrate in the following chapters, most of these works only consider one data modality or use sub-optimal network architectures and training strategies for the available data.

Therefore, the motivation for this thesis is to highlight and overcome the issues that have impeded operational FHM using multi-modal drone imagery, specifically RGB and thermal modalities. This is achieved by proposing new methods for the preprocessing and automated analysis of drone imagery using ML and DL-based computer vision methods. Specifically, I propose (1) an orthomosaicking workflow for RGB-thermal drone images of forests, which can potentially be used for other modalities like RGB-multispectral or RGB-hyperspectral, (2) a DL-based ITCD model that uses both RGB and thermal images for more precise detection results than using either modality in isolation, and (3) a DL-based classification model for distinguishing the severity levels of bark beetle-induced tree mortality without the need for a large-scale dataset. Together, these works form a complete processing pipeline that forest managers can employ for the FHM bark beetle infestation mapping task. Each of the three works presented in this thesis can also be used independently of the others, which can benefit several other forest monitoring applications. For instance, the orthomosaicking workflow could be used for forest fire monitoring without the need to extract individual tree crowns. On the other hand, the ITCD model can be applied to individual drone images when orthomosaic generation is not required.

1.2 Thesis Statement

This thesis is one of the first to build a comprehensive workflow pipeline comprising ML/DL models for the specific FHM applications of thermal orthomosaicking, individual tree crown detection, and bark beetle attack stage classification using drone-based multi-modal RS data. I support my arguments by conducting separate experiments for these three applications, proposing a novel method for each that outperforms existing techniques. The novelty of the thesis statement is embedded in the novelty of these proposed methods. Three subsidiary statements based on each of the applications are set forth with the aim of verifying them,

1. Orthomosaicking of thermal images of forests can be accomplished by leveraging the intermediate SfM results of higher-resolution, simultaneously acquired RGB images and co-registering image pairs using a learned transformation matrix.

2. Tree crowns, especially those of shorter trees hidden in shadows, can be effectively identified by a combined RGB-thermal DL detection model that is initialized with RGB pre-trained weights and fine-tuned for thermal images without annotations.

3. Visible stages of bark beetle attacks can be distinguished using a deep classification model through transfer learning with limited training data.

1.3 Contributions

Within this thesis, I seek to validate my thesis statement and its subsidiaries by making novel contributions in three different areas – orthomosaicking with image registration, object detection, and image classification – as summarized in Fig. 1.1. First, I propose a new workflow for generating high-quality thermal orthomosaics from simultaneously acquired RGB-thermal drone images. Forest monitoring applications often require georeferenced information in the form of large-scale orthomosaics, created by undistorting and stitching overlapping nadir (i.e., downward facing) images captured

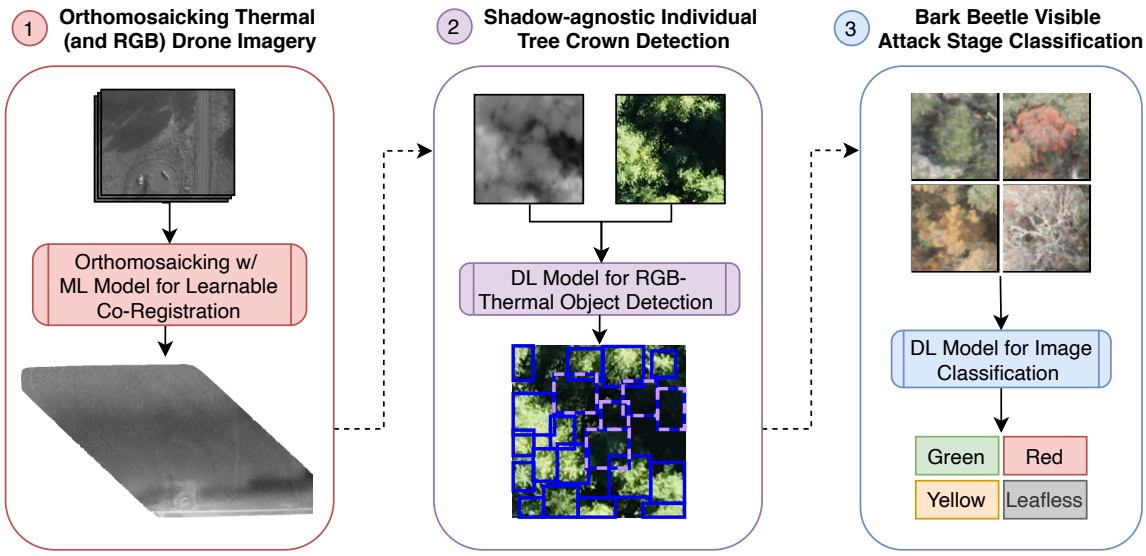


Figure 1.1: **Overview of Thesis Objectives** that collectively constitute a complete RS pipeline. Novel contributions are set forth for each objective in this thesis, as presented in the central chapters.

by drones. RGB cameras are commonly fitted on drones for low-cost, high-resolution imaging conducive to effective orthomosaicking, but only capture visible light. On the other hand, thermal sensors capture long-wave infrared radiation, which is useful for applications such as early pest detection or forest fire monitoring. However, these lower-resolution images suffer from reduced contrast and lack of descriptive features for successful orthomosaicking, leading to gaps or swirling artifacts in the orthomosaic generated by existing workflows. The proposed integrated orthomosaicking workflow overcomes these issues by using RGB images for producing a surface mesh via structure from motion, while using thermal images only to texture this mesh and yield a thermal orthomosaic. Before texturing, the RGB-thermal image pairs are co-registered using an affine transformation derived from a machine learning (ML) technique. On average, the individual RGB and thermal images achieve mutual information of 0.2787 on average after co-registration using the proposed technique, compared to 0.0591 before co-registration and 0.1934 using manual co-registration. In this thesis, I show that the thermal orthomosaic generated from my workflow (1) is of better quality than other existing methods, (2) is geometrically aligned with the RGB orthomosaic, (3) preserves

radiometric information (i.e., surface temperatures) from the original thermal imagery, and (4) enables easy transfer of downstream tasks—such as tree crown detection from the RGB to the thermal orthomosaic. Furthermore, I developed an open-source tool that implements the proposed workflow to facilitate usage and further development[1].

Following the orthomosaicking workflow, I present a novel deep learning (DL) model that successfully detects individual tree crowns from overhead forest imagery. This task poses a significant challenge due to the dense nature of forest canopies and diverse environmental variations, e.g., overlapping crowns, occlusions, and varying lighting conditions. The lack of data for training robust models also adds another limitation in effectively studying complex forest conditions. To tackle these problems, I (1) propose an entirely self-supervised method to effectively detect tree crowns despite challenging lighting conditions, and (2) present a challenging dataset comprising over 52k paired RGB-thermal images to facilitate future research for illumination-invariant detection. The proposed method leverages domain adversarial training to adapt from the RGB to the thermal data, using only the registered nature of image pairs as a supervision signal and thus bypassing the need for arduous manual annotations. During inference, extracted features from RGB and thermal modalities in poorly-illuminated areas are fused to effectively improve upon the predictions of an RGB-only detector and boost the overall precision and recall of detections. Through extensive experiments on the collected dataset, I demonstrate the proposed method's superiority over the baseline RGB-trained detector and previous state-of-the-art (SOTA) techniques that rely on unsupervised domain adaptation or early image fusion.

Finally, I propose a novel method to classify thermal tree crown patches into one of four visible bark beetle attack stages. Bark beetle outbreaks can dramatically impact forest ecosystems and services around the world. For the development of effective forest policies and management plans, the early detection of infested trees is essential. Over time, the crowns of infested pine trees turn from a healthy green to

[1]https://github.com/rudrakshkapil/Integrated-RGB-Thermal-orthomosaicing

yellow, then red, and finally grey (i.e., needleless). These are referred to as the attack stages. Despite the visual symptoms of bark beetle infestation, distinguishing between attack stages remains challenging, considering overlapping tree crowns in dense forest settings and non-homogeneity in crown foliage discoloration. The proposed method overcomes these challenges by leveraging the pre-trained feature extractor of a deep individual tree crown detector and using transfer learning to train a newly-introduced shallow subnetwork for classification. The experiments are performed on an existing bark beetle attack dataset originating in a forest stand in Northern Mexico. Various data augmentation strategies are examined to address the class imbalance problem in this dataset. Consequently, the affine transformation is selected to be the most effective one for this transfer learning task. Experimental evaluations demonstrate the effectiveness of the proposed method by achieving an average accuracy of 98.95%, considerably outperforming the existing baseline method by approximately 10%. The classification model along with code for training are publicly available[2].

Though these contributions target different research areas within the computer vision domain, together they encompass this thesis' main goal of solving practical RS problems for FHM. The works can be combined together into one RS pipeline that takes raw RGB-thermal drone data, generates a pair of orthomosaics, uses them to extract tree crowns, and classifies the bark beetle attack stage of these crowns.

1.4 Thesis Layout

This thesis is structured into five chapters, with the central chapters each comprising a part of the bigger picture and covering one aspect of my overarching goal — effectively applying computer vision techniques to advance FHM. Fitting all these parts together results in a comprehensive pipeline for RS applications that leverages deep learning during each substantial step. Following the introduction in Chapter 1, an improved method for generating high-quality orthomosaics from thermal images is proposed in

[2]https://github.com/rudrakshkapil/BarkBeetle-Damage-Classification-DL

Chapter 2. Orthomosaicking is typically one of the first preprocessing steps in drone-based RS applications. Then, a novel self-supervised strategy to detect individual tree crowns from overhead drone imagery is proposed in Chapter 3. In Chapter 4, extracted tree crown patches are classified into one of four bark beetle attack stages using a proposed transfer learning technique. Finally, a summary of the contributions of this thesis followed by a holistic discussion on the main takeaways, limitations, and future research directions is provided in Chapter 5.

Chapter 2

RGB-Thermal Orthomosaicking Pipeline

2.1 Introduction and Related Work

Forests are essential ecosystems that provide immense economic, social, and ecological value. Monitoring forest health is critical to understanding these ecosystems' challenges and devising successful management strategies to foster healthy and resilient forests [104]. For example, forest monitoring is important for detecting bark beetle attacks [37, 83] and assessing tree losses due to deforestation [98]. In recent years, drones have become increasingly popular for close-range monitoring of forests to capture high-resolution images of focus areas [20, 21, 52, 81]. Drones can be fitted with multiple-sensor instruments that take synchronized nadir images at regular intervals as the drone flies over the area to be imaged [114]. Optical sensors that capture information from various parts of the electromagnetic spectrum are common for forest health monitoring applications [21]. Among them, RGB sensors take images in the visible range of the spectrum (i.e., three channels – red, green, and blue) and have been extensively used for many applications, owing to their low cost and high resolution [30]. On the other hand, thermal sensors capture images that measure surface temperatures, which is beneficial for numerous applications, e.g., forest fire monitoring [87] or detection of insect-induced canopy temperature increases [118], for instance, those created by bark beetle infestations of Norway spruce (*Pinus abies*)

trees [147]. To facilitate monitoring in Geographic Information Systems (GIS) for such applications, drone-collected images must be orthorectified and stitched together to generate an orthomosaic. For instance, a GIS polygon layer with the location, size and shape of the crowns of young trees can be automatically derived by a predictive model from a drone RGB orthomosaic of a regenerating cut block [15].

Existing workflows typically generate orthomosaics for RGB and thermal data of forests separately [48]. However, thermal images typically lack enough contrast and salient features to enable smooth orthomosaicking [42], and temperature fluctuations can create different salient features in overlapping images from adjacent flight lines [79]; hence some practitioners prefer to skip the thermal orthomosaic entirely and directly work with individual thermal images [80, 118], but this complicates the transferal of information to GIS. Specifically, current standard methods that work well for RGB images [30] cannot reliably generate orthomosaics when it comes to thermal images of forests and crop fields [19, 44, 79, 105]. Thermal orthomosaics generated with these standard workflows suffer from swirling artifacts and gaps, leading to incomplete coverage of the entire area of interest (Fig. 2.1a). These artifacts are caused by poor depth estimation during the structure from motion (SfM) [128] stage of orthomosaicking [138].

Previous works have attempted to leverage simultaneously acquired RGB imagery to overcome the drawbacks of thermal-only workflows. A technique to improve the thermal orthomosaicking workflow by deriving the thermal image positions using RGB image alignment was proposed in [79]. Once the RGB image alignment is optimized, the external orientation parameters (*xyz* coordinates along with pitch, yaw, and roll angles) of each image are transferred to its thermal counterpart, and these are used as initial parameters in the SfM workflow for the set of thermal images. While a viable option, the second SfM process may introduce artifacts due to the previously mentioned issues inherent to thermal images. Likewise, [85] performed SfM for RGB and thermal images separately before registering (aligning) the resulting point clouds

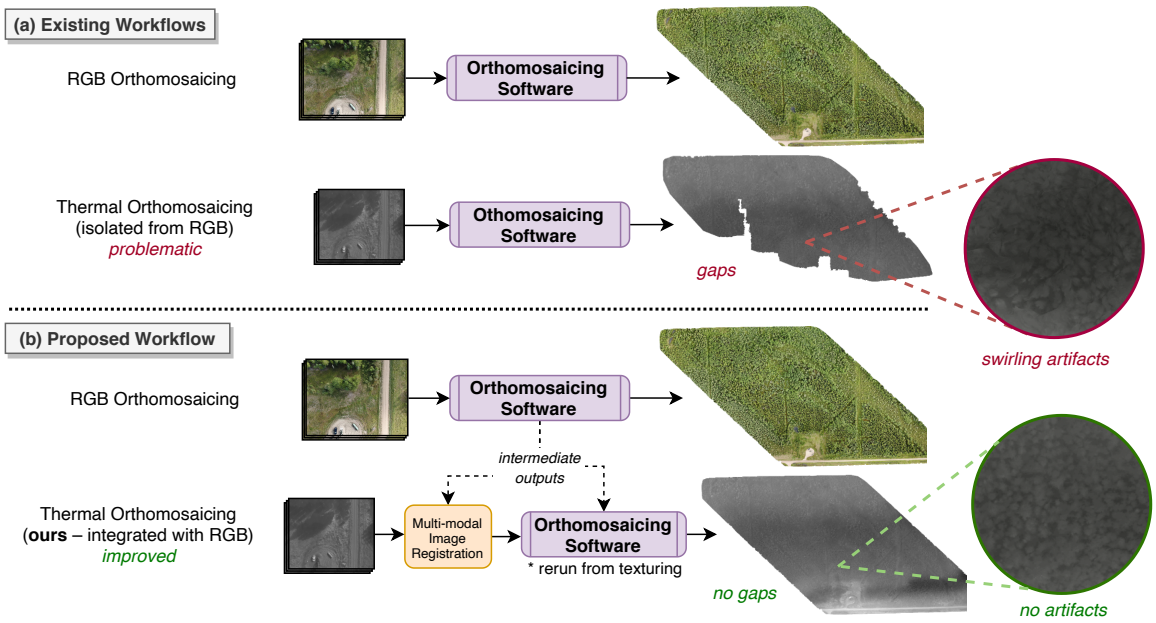


Figure 2.1: Comparison of **(a) Existing Workflows** for thermal orthomosaic generation and **(b) The Proposed Workflow** that leverages intermediate outputs from RGB orthomosaic generation. Thermal-only processing workflows are prone to gaps and swirling artifacts (shown in red), which are tackled by the proposed workflow.

for multi-modal 3D reconstruction of buildings. However, the thermal SfM step can introduce issues here as well. Similar to [79], [117] proposed a workflow in which the external orientations of the RGB-thermal image pairs were aligned but without any subsequent registration step. As the authors observed, using unregistered images leads to errors in the mosaic, even for the urban setting they considered. Another combined approach was proposed in [144], where each pair of RGB and thermal images was stacked into a 4-channel image prior to orthomosaicking and separated afterwards. However, this method requires an additional object-based geometric alignment stage that relies on the manual selection of clearly visible objects in both types of images, and that cannot fully resolve distortions due to different focal lengths in the lenses of the thermal and RGB cameras. Similarly, [49] used a technique that relies on manually-supplied pixel location correspondences between image pairs to fuse RGB and thermal point clouds of 3D structures. As precursors of the proposed automated co-registration (i.e., aligning the pixel-wise geometry of image pairs), the enhanced

correlation coefficient (ECC) [23] was applied with varying levels of success to RGB and thermal images of crop fields at different heights: [17] used ECC to register images taken 1.6m above the crop canopy, [72] applied ECC-based registration to drone imagery at a flight height of 30m, and more recently [73] used ECC to register images of olive groves at 45-50m altitude for generating aligned point clouds. However, these studies did not include the generation of orthomosaics. An edge feature-based image registration technique relying on image metadata was proposed in [63], whereas other works relied on point feature extraction [112, 126, 143]. Although these methods perform well for urban/generic settings, their application to nadir forest images is limited due to the lack of distinctive features in these scenarios, and moreover, they did not include orthomosaicking.

I propose a new integrated RGB and thermal processing workflow to overcome the mentioned challenges of thermal orthomosaicking of forest scenes. It is applicable to drone instruments that can simultaneously capture RGB and thermal images. The proposed workflow is summarized in Fig. 2.1b. It relies on an orthomosaicking algorithm based on texture mapping, as implemented in [97]. Texturing is a crucial step in this implementation of orthomosaicking, where rather than stitching together the individual orthorectified drone photos, a 2.5D mesh representing the outer envelope of the dense point cloud is created, and then the mesh is textured by projecting onto each small triangular mesh surface a particular patch of a drone photo from which it is best observed; the orthomosaic is then simply the orthographic projection of that textured mesh. Following this process applied directly to the RGB images, an orthomosaic is obtained along with important intermediate outputs – a surface mesh of the study site, the estimated external camera orientations used for texturing, and undistorted RGB images that have been corrected for radial and tangential distortions. Next, the simultaneously acquired RGB-thermal image pairs are co-registered, and the orthomosaicking process is rerun from the texturing step using the intermediate RGB outputs. This involves re-texturing the previously constructed surface mesh with the

thermal images based on the camera orientations estimated using the RGB images. Thus, the workflow generates two geometrically-aligned orthomosaics, where the same objects appear in the same pixel locations of both the RGB and thermal orthomosaics. Because thermal images are not used for SfM, which would perform poorly due to the low contrast and lack of features in these images, the gap and swirling issues present in thermal-only orthomosaicing workflows are successfully avoided, as shown in Fig. 2.1b.

Before the RGB intermediate outputs can be reused with the thermal images as described, the geometry of the individual thermal and RGB images needs to be precisely aligned to ensure that objects appear in the same pixel locations in each RGB-thermal pair. This is done through an intensity-based image co-registration method using gradient descent, inspired by [92] and [34]. In particular, the precise co-registration of RGB-thermal pairs is achieved through a multi-scale framework utilizing the matrix exponential representation [2, 35] and the normalized gradient fields (NGF) loss function [34]. According to [34] and confirmed by the results in Section 2.3.1, this loss function is more suitable than ECC for multi-modal drone imagery of forests. The co-registered thermal images replace the RGB undistorted images before rerunning the orthomosaicking process.

Rather than relying on manually selected objects like in [49, 144], the geometric alignment in the proposed workflow is automated and offers more degrees of freedom than just displacement. In addition, bypassing the SfM process for the thermal images helps overcome the issues of lower contrast and lower resolution and is thus preferable to previous feature-based techniques [63, 112, 126, 143]. The proposed workflow assumes that the RGB and thermal nadir images are captured simultaneously during the same flight (i.e., using a multi-sensor setup), which is the case for many commercially available drone cameras, for instance, the DJI Zenmuse H20T, DJI Zenmuse XT2, DJI Mavic 3T, senseFly Duet T, FLIR Hadron RGB/Thermal Camera Module, and Autel EVO II Dual 640T. A slight delay in the capture times of the two sensors is not problematic so long as the delay is systematic and an equal number of

RGB and thermal images are taken during the flight.

In this chapter, I demonstrate the effectiveness of my proposed orthomosaicking workflow using drone data from multiple dates over a forest stand in central Alberta, Canada. Aside from the high quality of the thermal orthomosaics generated (no gaps or swirling artifacts), I also verify that the orthomosaicking process preserves the radiometry of the images - both the individual thermal images and the corresponding part of the generated orthomosaic have the same thermal information. To highlight the utility of the proposed workflow that outputs geometrically-aligned RGB and thermal orthomosaics, I use a pre-trained DL model for individual tree crown detection (ITCD) from RGB images [136]. I show that the bounding boxes detected from the RGB orthomosaic can be directly used to extract tree crowns from the thermal orthomosaic, since they appear at the same pixel locations. As a further contribution, I provide a tool with an extensive graphical user interface (GUI) that implements the proposed RGB-thermal orthomosaic generation workflow. The developed tool is open-source to facilitate modification for specific projects and encourage additional functionality integration.

In summary, the contributions presented in this chapter are the following,

1. I propose an integrated RGB and thermal orthomosaic generation workflow that bypasses the need for thermal SfM by leveraging intermediate RGB orthomosaicking outputs and co-registering RGB and thermal images through an automated intensity-based technique.

2. I show that the proposed workflow overcomes common issues associated with thermal-only orthomosaicking workflows while preserving the thermal imagery's radiometric information (absolute temperature values).

3. I demonstrate the effectiveness of the geometrically-aligned orthomosaics generated from the workflow by utilizing an existing DL-based tree crown detector, showing how the RGB-detected bounding boxes can be directly applied to the

thermal orthomosaic to extract thermal tree crowns.

4. I develop an open-source tool with a GUI that implements the workflow to aid practitioners.

The rest of this chapter is organized as follows. Section 2.2 provides an overview of the study site for this work, followed by a detailed description of the stages comprising the proposed integrated RGB-thermal orthomosaicking workflow. That section additionally contains a description of the developed GUI tool. Section 2.3 presents quantitative and qualitative results of the orthomosaic generation using drone data from the study site, including the performance of the downstream ITCD task on the generated orthomosaics from the proposed workflow. In Section 2.4, I present a detailed discussion of the results and provide important recommendations for using the proposed workflow. Finally, I discuss possibilities for future work and present conclusions in Section 2.5.

2.2 Materials and Proposed Workflow

In this section, I first describe the dataset used for the experiments within this chapter. Then, I provide an in-depth description of the proposed integrated orthomosaicking workflow for thermal images. Following this, I explain how an example downstream task can be used to highlight the utility of the workflow. Next, I outline the developed GUI tool and provide recommendations for its effective use. In the end, I describe how I empirically assessed the performance of the proposed workflow.

2.2.1 Cynthia Cutblock Study Site

Drone data was repeatedly collected from an 8-hectare forest stand approximately 3.5km northeast of Cynthia, Alberta (Canada), called the Cynthia cutblock. The location is shown in Fig. 2.2. The region is at an elevation of around 950m above sea level. Lodgepole pine (*Pinus contorta ssp. latifolia*) and aspen (*Populus tremuloides*)

make up the majority of the tree species found within the cutblock. A Zenmuse H20T instrument mounted on a DJI Matrice 300 RTK quadcopter was used to take nadir RGB and thermal images of the cutblock. The H20T is fitted with three cameras. The wide-angle RGB camera takes images of 3040×4056 pixels and covers the most terrain (83° field of view (FOV)). The thermal images are 650×512 pixels and have a FOV of 41°. The zoom camera (not used in this study) can go up to 4° FOV and has dimensions of 5184×3888 pixels. The two RGB sensors are CMOS sensors, whereas the thermal one is an uncooled VOx microbolometer that produces 16-bit radiometric JPEGs (RJPEGs). This work considered five flights on different days in 2022 over the Cynthia cutblock using the described camera setup (July 20, July 26, August 9, August 17, and August 30). Each flight lasted approximately 30 minutes between 10 am and 1 pm. Weather conditions varied across all flights – air temperature was between $20°C$ and $25°C$, relative humidity was between 40% and 61%, and cloud cover was limited. Specific air temperature and relative humidity values during each flight are reported in Table 2.1, where the values are averages from the three weather stations closest to the cutblock. The drone was flown 120 meters above the ground, following the same flight path across all dates. The thermal and wide-angle RGB images were taken simultaneously, such that thermal images had a 75% side overlap and 80% front overlap. Approximately 800 image pairs were captured during all flights. The EXIF header of each JPEG image contains the GPS coordinates of the drone when that image was taken. This information is helpful for individual image localization and georeferencing during the orthomosaicking process. Each output orthomosaic covers an area of around 30 hectares encompassing the cutblock.

2.2.2 Proposed Integrated Orthomosaicking Workflow

Here, I describe the proposed workflow for generating a thermal orthomosaic by leveraging RGB data. The workflow relies on Open Drone Map (ODM) [97], an open-source framework that implements a texturing-based orthomosaicking algorithm

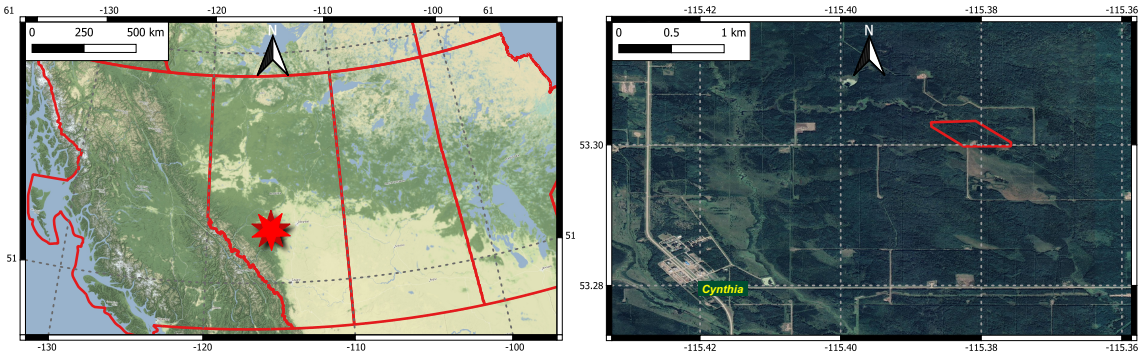


Figure 2.2: **Left**: Location of Cynthia cutblock in Alberta, Canada. **Right**: A close-up of the area including the village of Cynthia.

Table 2.1: **Summary of Cynthia Cutblock Data for each Flight Date**. Temperature and humidity values are taken from three weather stations closest to the cutblock and averaged.

	Jul 20	**Jul 26**	**Aug 09**	**Aug 17**	**Aug 30**
Number of RGB-Thermal Image Pairs	827	828	820	825	814
Average Air Temperature ($°C$)	20.3	20.8	19.8	24.5	25.4
Average Relative Humidity (%)	42.7	61.0	53.0	40.7	46.3

similar to that described in [40]. ODM is described in further detail in Section 2.2.2, but briefly, a textured surface mesh reconstruction of the scene is first produced to yield an orthomosaic, rather than relying on a digital surface model (DSM) for orthorectification of images before stitching. The intermediate outputs generated from the RGB orthomosaicking process (i.e., surface mesh reconstruction and external camera orientations) are used to initialize the thermal orthomosaic generation, bypassing the need for SfM with thermal images and therefore avoiding the issues present in thermal-only orthomosaicking workflows. Instead, thermal images are only used to texture the surface mesh previously reconstructed from the RGB images, thereby producing a high-quality thermal orthomosaic.

The proposed integrated orthomosaicking workflow comprises four stages, as shown in Fig. 2.3 – (1) RGB orthomosaic generation, (2) Thermal image conversion (from R-JPEG to grayscale TIFF), (3) RGB-thermal image co-registration, and (4) Thermal

orthomosaic generation. The proposed workflow has also been implemented as an open-source tool, presented with more details in Section 2.2.4.

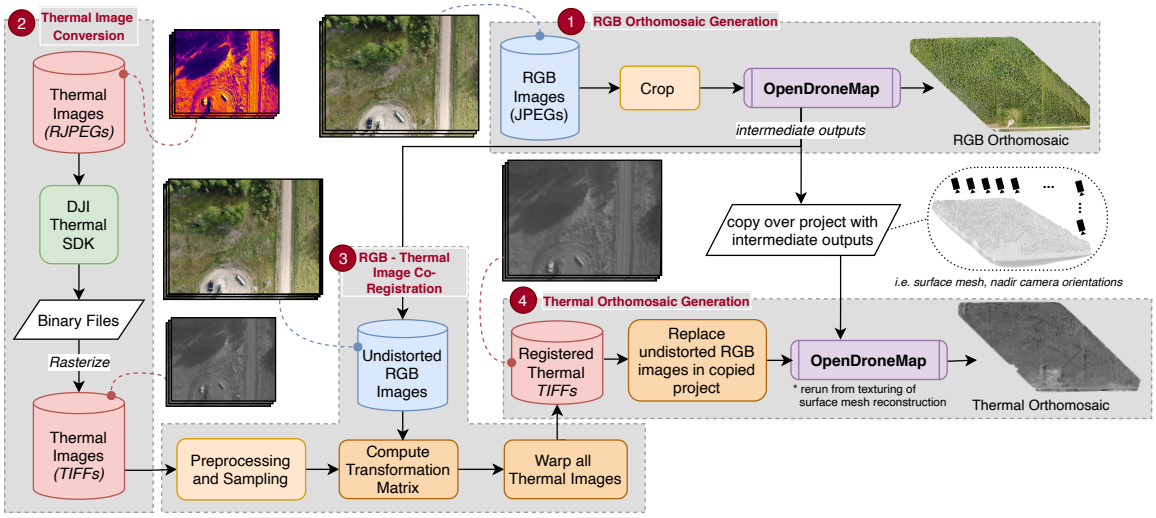


Figure 2.3: **Overview of the Integrated RGB-thermal Orthomosaicking Workflow**. Example images of each type of data are also included, connected by dashed lines. Stage 3 is more detailed in Fig. 2.4.

The rest of this section provides a summary of the integrated workflow, and the following subsections describe each stage in more detail. In the first stage, ODM is used to generate an RGB orthomosaic. This relies on structure from motion (SfM) [128], a computer vision process that automatically aligns the overlapping drone photos. After this, a dense point cloud and a surface mesh of the study area are derived together with the estimated camera orientations, as shown in the dotted bubble on the right side of Fig. 2.3. Undistorted RGB images are also obtained as intermediate outputs – these are the original input RGB images but corrected for radial and tangential distortions. After the RGB orthomosaic is generated, the ODM project is duplicated along with its intermediate outputs.

In the second stage, each thermal image is preprocessed to obtain, out of the native R-JPEG format, a single-channel (i.e., grayscale) image that displays the surface temperature of the object present within it. The third stage is the computation of the transformation matrix that co-registers every preprocessed thermal image with its

corresponding undistorted RGB image, such that both are geometrically aligned and objects appear in the same pixel locations. Note that this geometric alignment (i.e., co-registration) is different from the 'alignment' during SfM, which yields the external orientations of the cameras.

Finally, in the fourth stage, the undistorted RGB images in the duplicated project are replaced with the co-registered thermal images obtained from the previous stage, and the ODM process is restarted from the texturing step. In this way, SfM with thermal images is avoided. Instead, the co-registered thermal images are used to texture the surface mesh previously derived from the RGB images. This stage outputs a high-quality thermal orthomosaic that is geometrically aligned with the RGB orthomosaic.

RGB Orthomosaic Generation

In the first stage of the proposed workflow, ODM is used to generate the RGB orthomosaic of the study site from the collected RGB nadir drone images. If the FOVs of the RGB and thermal cameras differ drastically, such as for the H20T instrument, the central regions of the RGB images are cropped so that each RGB-thermal pair depicts roughly the same scene. This is necessary for effective geometric alignment in the RGB-thermal image co-registration stage later on. The exact amount of cropping depends on the specific cameras used. The cropping must generally maintain sufficient forward and sideways overlap between adjacent images for effective orthomosaicking. In this case, the 3040×4056 wide-angle RGB images were cropped to 1622×1216 pixels (60% reduction in width and height). An additional advantage of cropping wide-angle RGB images is that it mitigates lens distortion from the edges of such images from propagating to the orthomosaic. Without cropping, relief displacement (tree lean) effects would likely occur in the RGB orthomosaic due to the relatively low flight height [138]. Moreover, reducing the image size by cropping enables faster processing in ODM. Once the RGB orthomosaic is successfully generated by ODM, the project

is duplicated for reuse in later stages.

The key steps involved in the orthomosaic generation process of ODM are summarized as follows. The 3D structure of the area of interest (AOI) is reconstructed using the SfM technique within ODM; namely, OpenSfM [82]. This involves extracting tie points (corresponding to salient features) from images using the scale-invariant feature transform algorithm [75] and then matching tie points from overlapping images. These points are then located in 3D space using parallax. Next, the external orientation of the camera for each image is refined in an iterative process called bundle adjustment. Finally, new points are added to the initial point cloud of tie points using multi-view 3D reconstruction based on depth estimation [115], yielding a dense point cloud with thousands of points per square meter. Thus, the 3D structure of the scene is inferred entirely from (RGB) drone imagery without the need for expensive sensors like LiDAR.

The SfM process additionally performs internal camera calibration, providing estimates for the camera model's focal length, principal point, and distortion coefficients. The same internal camera parameters apply to all RGB images since they are all captured by the same sensor. The estimated distortion coefficients are used to undistort every RGB image under the Brown-Conrady distortion model [12]. This models the radial distortion in the RGB images using coefficients k_1, k_2, and k_3, and tangential distortion with coefficients p_1 and p_2, both arising from the camera lens shape. Briefly, every pixel location in the original image is mapped to its corresponding location in the undistorted image, correcting for the distortions using these coefficients. More details about the exact formulation can be found in [12]. Compared to the (distorted) image taken from the wide-angle lens, the undistorted image is warped to look closer to what it would look like using a pinhole camera. For instance, straight lines in the real world that appear curved in the captured image are recovered as straight in the undistorted image.

OpenMVS [14] is then used to produce a textured mesh with the undistorted images: outliers in the dense point cloud are filtered out, and the 2.5-dimensional surface

mesh, i.e., a warped surface typically made of small triangles that represents the outer envelope of the point cloud, is generated using Screened Poisson Reconstruction [55]. At this point, the estimated external orientations (one for each image) and camera internal parameters (same for all images) are used to determine how to texture the triangular mesh surfaces, which involves selecting for each small triangle the undistorted image from which it is best observed. This is done through Triangle to Image Assignment, which considers each image's proximity to the triangle, the triangle's amount of occlusion (if any), and the viewing angle's steepness relative to the triangle surface [40]. Once assigned, the relevant pixels of the undistorted images are projected onto the triangular surfaces using the calculated transformation matrices from the exterior and interior parameters. Finally, the orthomosaic is generated as the orthographic projection of the textured mesh onto the horizontal plane defined by the chosen coordinate system (*NAD83 UTM 12 N* in this case). The default ground sampling distance (GSD) corresponds to the average pixel size on the mesh surface. This results in a single, high-resolution RGB orthomosaic of the study area. This texturing-based orthomosaicking differs from alternative methods that use a digital surface model to orthorectify individual images that are later stitched together.

Thermal Image Conversion

In the next stage of the proposed workflow, the 3-channel, 8-bit RJPEG images captured by the H20T camera, which are intended for visualization (as in the top left of Fig. 2.3), are converted to absolute surface temperature readings in Celsius degrees, stored as 32-bit floating-point TIFF images. This is done through the DJI Thermal SDK (software development kit) using the 'measure' functionality, with average emissivity as 0.95 and distance to target as ¿25m for all flights. Relative humidity and reflected temperature (i.e., air temperature) are set independently for each flight using the values reported in Table 2.1. The SDK outputs one binary file per image, which is then rasterized into grayscale TIFF images like those in Fig. 2.3.

RGB-Thermal Image Co-registration

As a result of the initial cropping of the RGB images, the footprint of a preprocessed thermal image roughly coincides with that of its corresponding undistorted RGB image. In this stage, the geometry of the image pairs is more precisely aligned through image co-registration. Specifically, this requires the computation an estimate to the optimal geometric transformation matrix M^* that co-registers every thermal image with its corresponding undistorted RGB image, such that objects appear in the same pixel locations in both. In this work, the computation is restricted to a single affine linear transformation matrix, i.e., a 3×3 matrix with 6 degrees of freedom that preserves parallel lines. The same transformation applies to all image pairs as they come from the same instrument simultaneously capturing both modalities. This has the advantage of being computationally efficient while not adversely impacting co-registration performance, as corroborated in Section 2.3. Once a close approximation to the optimal M^* is obtained, it is used to warp all the thermal images to emulate their undistorted RGB counterpart, using the concept of inverse warping [123]: For every pixel location (x_{out}, y_{out}) in the warped output thermal image, the pixel location (x_{in}, y_{in}) in the input thermal image is calculated by performing matrix multiplication of the inverse of M^* with the homogenous coordinates $(x_{out}, y_{out}, 1)^T$. The third dimension is used to scale coordinates in the projective plane and set by convention to 1. Then, the value at the pixel location (x_{in}, y_{in}) in the unwarped image is assigned to (x_{out}, y_{out}) in the warped image, performing re-sampling through bicubic interpolation.

This work investigates two methods for computing the optimal transformation matrix M^*. One is to manually supply up to four point correspondences between a pair of RGB and thermal images and then to solve a system of linear equations for an estimate of M^* [41]. Although these correspondences need only be supplied once, the quality of co-registration (and optimality of the computed transformation matrix) heavily depends on their correctness.

In the second method, the estimate of M^* is automatically computed through intensity-based image co-registration using gradient descent optimization, inspired by [92] and [34]. Given a set of N thermal images and their corresponding N RGB counterparts, the gradient descent optimization iteratively refines the 6 learnable parameters (i.e., variables) within a Homography Module that encapsulates the computation of the transformation matrix M [92]. An intensity-based co-registration is leveraged, which does not rely on extracting features from both images, instead working directly with the pixel values and image gradients [34]. Fig. 2.4 summarizes this process, and a detailed description follows below.

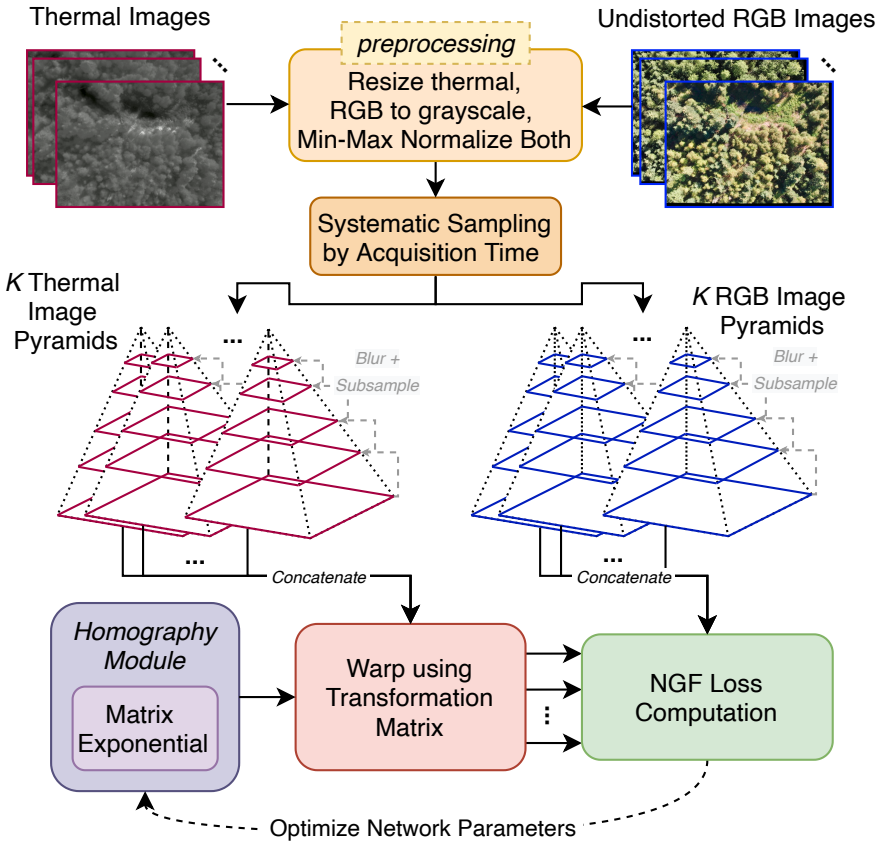


Figure 2.4: **Detailed Steps of the Automated Intensity-based Image Co-registration**. After preprocessing the images, K RGB and thermal image pairs are systematically sampled for batch processing. The parameters of the Homography Module [92] represent the values in the transformation matrix being computed and are learned during gradient descent optimization [106] of the normalized gradient fields (NGF) loss [60] between each transformed thermal Gaussian pyramid and its corresponding RGB pyramid.

First, both sets of images need to be preprocessed to facilitate further computations. The grayscale thermal images extracted during the previous stage are upscaled through bicubic interpolation to match the dimensions of the undistorted RGB images (i.e., from 640×512 to 1622×1216). The resized thermal images are then min-max normalized. As for the undistorted RGB images, they are converted to their single-channel luminance representation using the weighted formula $L_{(x,y)} = 0.2125R + 0.7154G + 0.0721B$ for each pixel location (x, y), as defined in the *rgb2gray* function of the scikit-image processing library [131]. The resulting grayscale images are also min-max normalize.

The gradient descent-based optimization is a computationally intensive process. To avoid loading all image pairs into the hardware memory (CPU or GPU) while still ensuring the transformation matrix computation is robust, a batch of 64 image pairs is sampled. 64 was deemed sufficient after empirically testing different batch sizes. Further, working with a batch rather than a single pair ensures less variance in gradient calculations while encouraging good convergence during gradient descent [106]. Although any image pairs can be selected for the batch as a single linear transformation matrix that applies to all pairs needs to be computed, appropriately choosing batch pairs is important for good co-registration performance. Therefore, systematic sampling is performed: the RGB-thermal image pairs are first ordered by acquisition time, then every j-th pair is picked for the batch, where $j = \lfloor N/K \rfloor$ and K is the batch size. Besides the intuitive reasoning that the computed M should be more robust due to a greater coverage of the AOI within the batch, the quantitative results in Section 2.3.1 will also show that systematic sampling is preferable to random sampling for batch pairs. A multi-resolution Gaussian pyramid is then constructed for all the grayscaled, normalized RGB and thermal image pairs in the batch, where each smaller layer is obtained by blurring followed by sub-sampling from the previous larger one [2]. A batch size of 64 is utilized, with 11 levels in each pyramid at a downscale factor of 1.5. Blurring is done with a filter mask twice the size of the downscale factor that covers more than 99% of the Gaussian distribution, and sub-sampling is done

through pixel averaging. This multi-resolution framework ensures that the current estimate of M will be refined simultaneously at multiple scales, allowing for more precise and robust co-registration [78].

During gradient descent, the following objective function is optimized to find the transformation matrix M that warps every thermal image T to most closely align its geometry with that of its corresponding RGB image R [92],

$$\min_{M} L(R, Warp(T, M)), \tag{2.1}$$

where L is the normalized gradient fields (NGF) loss function [60]. This loss function is well suited for optimization and preferable for multi-modal images [34]. It is based on the principle that two images are well-registered if intensity changes occur at the same locations. NGF computation requires the x-direction gradient ($\nabla_{\mathbf{x}} I(x, y)$) and y-direction gradient ($\nabla_{\mathbf{y}} I(x, y)$) of the image I at every pixel location (x, y), which in this work is numerically approximated using central finite difference [139],

$$\nabla_{\mathbf{x}} I(x, y) \approx \frac{I(x+1, y) - I(x-1, y)}{2}, \qquad \nabla_{\mathbf{y}} I(x, y) \approx \frac{I(x, y+1) - I(x, y-1)}{2} \tag{2.2}$$

Then, the NGF loss $L(I, J)$ for two grayscale images I and J as [34],

$$L(I, J) = \frac{1}{w \cdot h} \sum_{x=1}^{w} \sum_{y=1}^{h} [(\frac{\nabla_{\mathbf{x}} I(x, y)}{\|\nabla I(x, y)\|} - \frac{\nabla_{\mathbf{x}} J(x, y)}{\|\nabla J(x, y)\|})^2 + (\frac{\nabla_{\mathbf{y}} I(x, y)}{\|\nabla I(x, y)\|} - \frac{\nabla_{\mathbf{y}} J(x, y)}{\|\nabla J(x, y)\|})^2], \tag{2.3}$$

where h and w are the number of rows and columns in both images and $\|\nabla I(x, y)\|$ denotes the point-wise gradient magnitude at pixel (x, y).

For the gradient descent-based optimization to work, M is encapsulated as the learnable parameters v within a simple differentiable module, termed the Homography Module [92], using the matrix exponential representation [36],

$$exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{2.4}$$

The matrix exponential formulation has the following two desirable properties [35]. Its computation is differentiable, which is necessary for learning the Homography

Module's parameters v through backpropagation during the gradient descent optimization. Moreover, the output of the matrix exponential function (i.e., the computed transformation matrix) is always invertible, so forward and inverse transforms can be reliably combined for more robust NGF loss computation. In practice, using 10 terms closely approximates the sum of the infinite series. Using the matrix exponential representation, the objective function in Equation (2.1) can be rewritten as,

$$\min_{v} L(R, Warp(T, M(B, v))), \tag{2.5}$$

$$M(B, v) = exp(\sum_{i=1}^{6} v_i B_i), \tag{2.6}$$

where $M(B, v)$ is the current estimate of the affine transformation matrix, derived from the current parameters $v = \{v_i | i = 1...6\}$ of the Homography Module and the constant basis matrices $B = \{B_i | i = 1...6\}$. These 3×3 matrices are generators of the group of affine transformations on the 2D plane, as described in [92]. Therefore, any affine transformation matrix may be computed using this formulation. Before commencing training, $v_i = 1 \forall i = 1, .., 6$ is initialized. Fig. 2.5 summarizes the mechanism of the Homography Module and how matrix exponential is implemented within it.

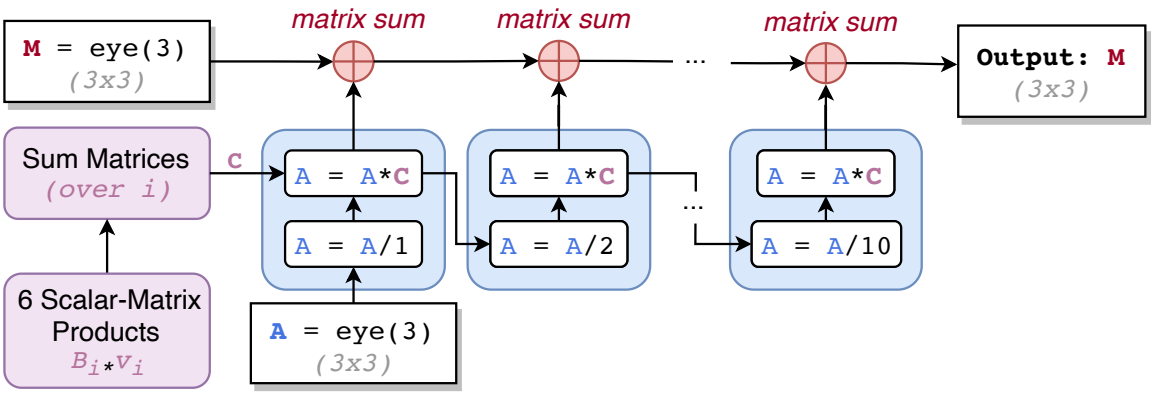


Figure 2.5: **Internal Mechanism of the Homography Module**, which leverages matrix exponential to compute the affine transformation matrix M using 6 basis matrices B and learnable scalar parameters v. The same 3×3 matrix C is used to update A at each of the 10 substeps, which are finally summed to yield M. The term *eye(3)* denotes the identity matrix of size 3×3.

At every step of the optimization loop in Fig. 2.4, the current parameters v of the Homography Module are used to derive M. Using M, the average NGF loss between RGB and warped thermal pairs is computed (forward transform loss) along with the loss between thermal and warped RGB pairs (inverse transform loss) at every image pyramid level, and both of these losses are summed. The total multi-resolution loss is the sum of bi-directional losses across all pyramid levels. Backpropagation [107] is used to learn the parameters within the Homography module v by using the partial derivatives with respect to v of the computed multi-resolution NGF loss to adjust v in a way that decreases loss. The Adam optimizer [59] (as implemented in the PyTorch software library [101]) is utilized with a learning rate of 0.005. All other optimizer hyperparameters are unchanged from their defaults. After each training iteration, the computed loss should decrease, and the estimates of v should be closer to their optimal values. Training is terminated once convergence is reached (i.e., loss stops decreasing), for which 200 iterations were found to be sufficient during all the experiments. The best approximation to the optimal transformation matrix M^* has been found at convergence. In the final step of this stage, the computed matrix is used to warp all the resized, non-normalized thermal images. The non-normalized images are warped to preserve the original absolute temperature values during orthomosaicking in the next stage.

Note that the described intensity-based co-registration differs from previous feature-based registration techniques that match salient feature descriptors between the images, for example in [63]. It is also different from the mutual information (MI)-based registration in [92], which trains an additional neural network to approximate MI for image pairs as a measure of similarity of their grey-level histogram distributions, and performs gradient descent with respect to MI loss to compute the transformation matrix for registration. To compute MI, they consider pixels belonging to edges (determined by Canny edge detection). However, in this work, all pixels are used to compute NGF loss since canny edge detection was observed to not be robust for

the aerial forest images in the Cynthia cutblock data. Lastly, in comparison to the pair-specific diffeomorphic (non-linear) transformations used in [34] for medical image registration that deforms images unevenly, the transformation in this work is restricted to a single linear transformation matrix for all image pairs, which is both efficient and effective for this work as the results corroborate in Section 2.3.

Thermal Orthomosaic Generation

In the final stage of the proposed workflow, the undistorted RGB images in the duplicate of the RGB orthomosaic project are first replaced with the co-registered thermal images obtained from the previous stage. Then, this duplicated ODM project is rerun, starting from the texturing step. The same external camera orientations estimated from the RGB SfM process during orthomosaicking are reused here. As a result, the co-registered thermal images are used to texture the surface mesh previously obtained as an intermediate output of the RGB orthomosaicking process. In this way, ODM outputs a thermal orthomosaic. Because the individual undistorted RGB and thermal images were co-registered in the previous stage, the RGB and thermal orthomosaics are also co-registered (i.e., geometrically aligned). Each tree appears at the same pixel locations in both orthomosaics. Both orthomosaics also contain the same georeferencing information; hence, they line up exactly when viewed using GIS software (see the next section).

2.2.3 Downstream ITCD Task

To highlight the versatility and utility of the proposed workflow, in Section 2.3.5 I will demonstrate how the generated orthomosaics can be used for a practical downstream task, specifically tree crown detection. I will use DeepForest [136], a pre-trained deep neural network, to delineate individual tree crowns from the RGB orthomosaic. This model was originally trained using RGB images, not thermal ones. However, since the proposed workflow generates geometrically-aligned orthomosaics, the results will show

that the same bounding boxes obtained from the RGB orthomosaic apply directly to the generated thermal orthomosaic.

The DeepForest model is based on the one-stage object detection RetinaNet [67] framework. Its backbone comprises a ResNet-50 [43] neural network that extracts multi-scale features from the image, and a feature pyramid network (FPN) [66] that combines semantically low-resolution features with low-level, high-resolution ones. Each level of the FPN feeds its computation to a regression head that locates bounding boxes in the image and to a classification head that outputs a confidence score for each box. This score denotes the model's confidence that a tree crown is contained within the predicted box.

2.2.4 Proposed Open-Source Tool

To facilitate the usage of the proposed workflow, I have created an open-source tool that offers both a command line interface and a graphical user interface (GUI) developed using the PyQt5 Python library. As shown in Fig. 2.6, every setting for each stage of the workflow can be easily customized from the GUI depending on the specific requirements of a project. The tool runs on the Windows operating system and leverages GPU acceleration if available. It offers the option to toggle each processing stage, including the downstream ITCD, and specify their relevant hyperparameters. For example, it is possible to specify the batch size, the number of pyramid levels, and the degrees of freedom in the co-registration stage. Additionally, there is an option to supply manual point correspondences for a chosen image pair or to compute the transformation automatically using the described intensity-based co-registration. Each of the Open Drone Map settings can likewise be modified directly using the GUI, and it also offers the option of RGB- or thermal-only processing. Practitioners may use this tool without having to look at any code.

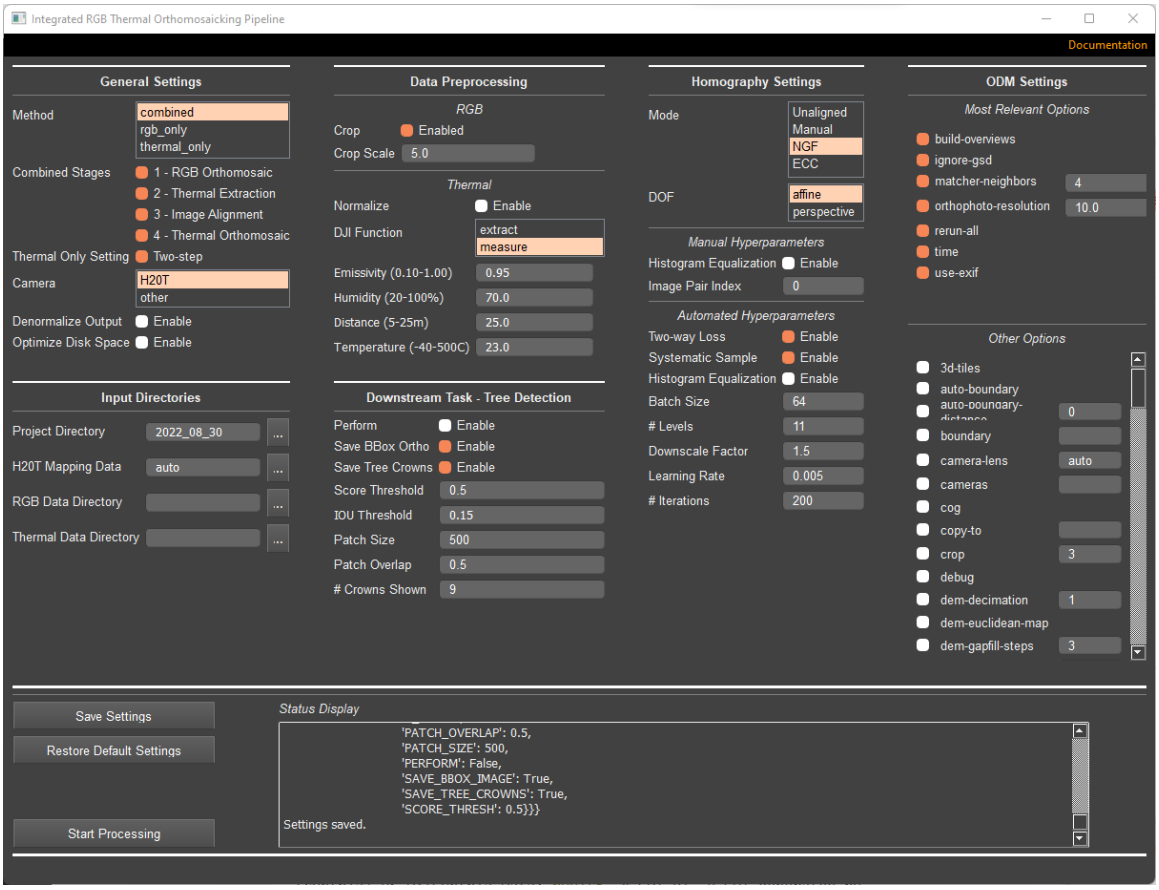


Figure 2.6: **Graphical User Interface** for the orthomosaicking tool showing the various easily-configurable processing options.

2.2.5 Performance Assessment

During the experimental evaluation for this work, all orthomosaics were generated at a ground sampling distance (GSD) of 10cm, which is close to the spatial resolution of the original thermal images at the nadir. The drone images were collected from the Cynthia cutblock, as described in Section 2.2.1. Although the same H20T camera was used during all flights, separate transformation matrices were computed for the geometric alignment of each flight individually. The quality of four different co-registration techniques on individual thermal and RGB image pairs were compared. The first was the baseline, i.e., not performing any co-registration of thermal and RGB images, similar to that in [117]. The second was manual co-registration, where an affine transformation matrix is computed using three point correspondences supplied

manually by a user when looking at a randomly chosen RGB-thermal pair, similar to that in [49]. The best result across 10 repeated trials has been reported. The third technique is the proposed intensity-based registration framework using the NGF loss function, as described in Section 2.2.2. For the fourth, the enhanced correlation coefficient (ECC) [23] was used as the loss function instead of NGF in the proposed workflow. ECC seeks to maximize the pixel-wise correlation between image pairs and has been used in previous works for RGB-thermal co-registration [17, 72, 73]. The four techniques were compared using mutual information (MI) as a quantitative metric, which measures how well the intensity in one image can be predicted given the intensity in the other. Thus, MI can be used to measure the similarity in grey-level distributions for two images from their histograms. MI was also used in further experiments to determine the optimal design choices for other parts of the proposed workflow, such as batch size and sampling method. Additionally, the co-registered images and orthomosaics for the proposed NGF-based workflow were qualitatively compared to the unregistered baseline. Then, the Bhattacharyya coefficient [8] was used to compare the histograms of randomly chosen original thermal images and their corresponding patches in the thermal orthomosaic to confirm that the proposed workflow preserves radiometric information in the form of absolute temperature values. Finally, the performance of the downstream ITCD task on the thermal orthomosaics using the RGB-detected tree crowns was evaluated. The observed results are discussed in depth in Section 2.4.

2.3 Experimental Results

2.3.1 Quantitative Results

Image registration techniques are typically evaluated against a 'gold standard' registration using a measure of positional error between known ground truth correspondence points when available [103]. However, this information was unavailable for the Cyn-

thia study site for the two modalities, as many of the image pairs lack any salient geometric feature to be used as ground truth. Specifically, points on swaying tree crowns due to wind are unsuitable, and easily identifiable hotspots in thermal images do not necessarily correspond to salient features in the RGB images (and vice versa). Therefore, the MI between an RGB and thermal image pair was utilized as a quantitative metric to measure the performance of each image co-registration technique. Better co-registration corresponds to a higher MI value. MI was computed using the histograms of the min-max normalized images with 100 equally spaced bins for all image pairs, as described in [92].

Fig. 2.7 shows the MI between the individual RGB and thermal images from the August 30 Cynthia flight. The unregistered images had a median MI of 0.054. Performing manual image co-registration improved the median MI of the image pairs to around 0.203. Although this is a considerable improvement, it could be even higher if more accurate point correspondences are provided. The transformation matrix computed automatically using intensity-based co-registration with NGF as the loss function resulted in the highest MI, with a median value of 0.324. Using ECC instead of NGF produced a significantly lower median value of 0.198, close to the performance achieved by manual registration. Except for July 26, where ECC performed slightly better than NGF, similar results in terms of relative performance were obtained for the other flights (Table 2.2). These results indicate that the intensity-based co-registration using NGF most closely aligned the geometry of the RGB and thermal images compared to the other co-registration techniques.

In Table 2.2, I report the average MI after co-registering all individual thermal and RGB images using different design choices within the proposed workflow. There is one column for each flight in the Cynthia cutblock data, and the final column reports the arithmetic mean over all five flights. The rows are divided into multiple sections, one for each design choice. Each design choice was tested independently while setting all others to their best-performing options (indicated in bold). I make

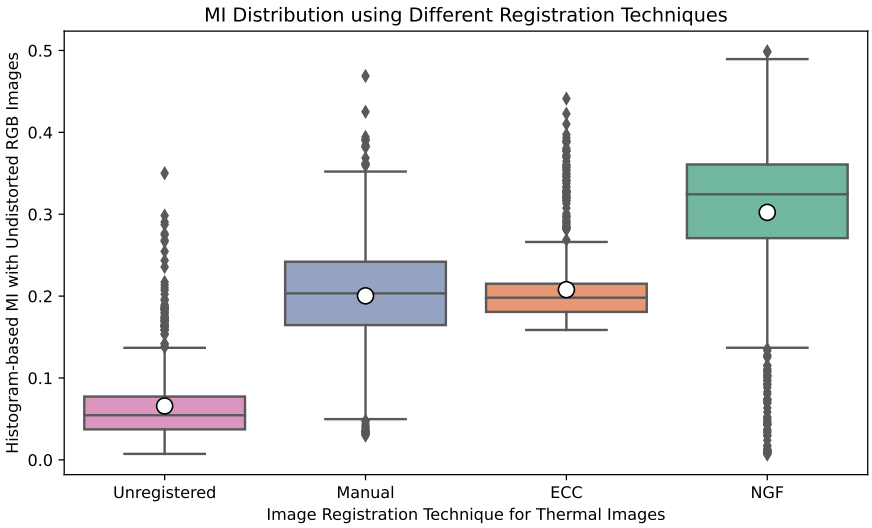


Figure 2.7: **Mutual Information (MI) for August 30 Cynthia Flight**. Box (interquartile range, IQR) and whisker (within 1.5× IQR) plots are shown for the MI between 814 RGB-thermal image pairs using different co-registration techniques. NGF is normalized gradient fields [34], while ECC is enhanced correlation coefficient [23]. The white circles denote mean values.

the following observations from the reported results: (1) Co-registration using NGF outperformed the other techniques on average across the five flights by 0.2196 units over unregistered, 0.0853 units over ECC, and 0.0575 units over manual; (2) Restricting the transformation matrix to six degrees of freedom (affine) resulted in a slightly higher average MI than allowing eight degrees of freedom (perspective); (3) The multi-resolution Gaussian pyramid framework performed immensely better than using only the single, highest-resolution image by 0.2519 units; (4) Larger batches led to a higher average MI, with a batch size of 64 yielding the most performance improvement by least 0.02 units over smaller batches; and (5) Systematically sampling image pairs for the batch marginally outperformed random sampling. Setting each design choice to its best-performing option yielded the highest average MI of 0.2787, indicated in bold. The average MI between the perfectly registered red and blue channels of the same RGB images from the August 30 flight was 1.4896, which is predictably higher than the reported values between grayscaled RGB and thermal pairs of different modalities.

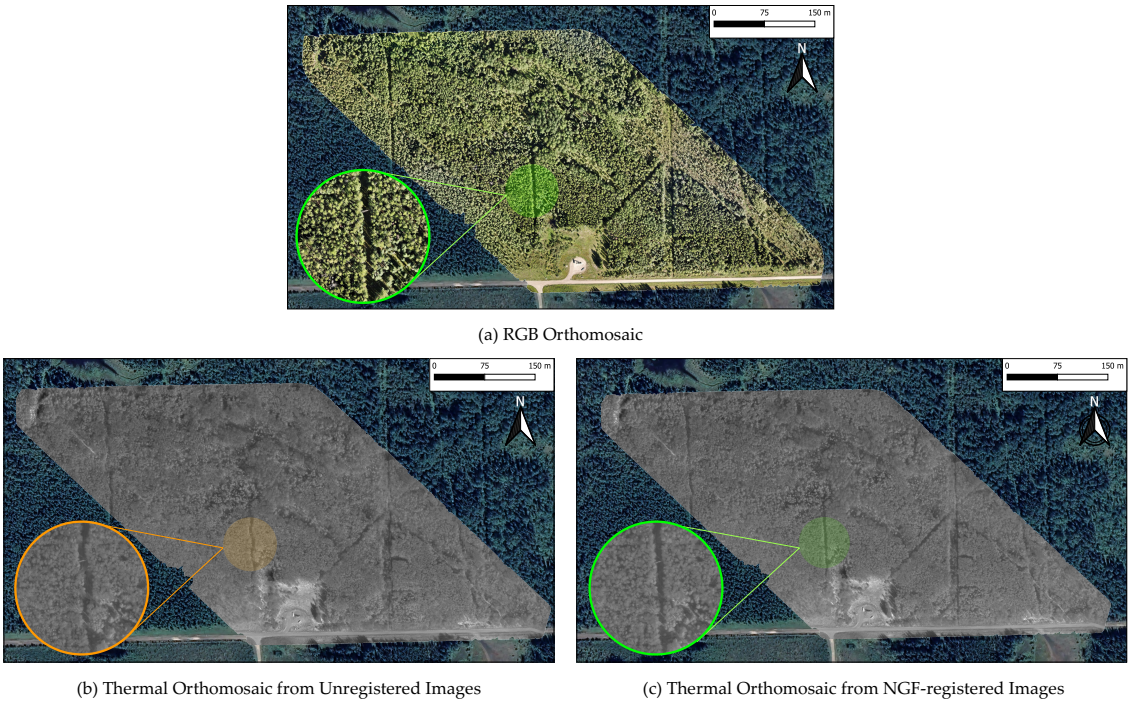
Table 2.2: **Average MI of Individual RGB and Thermal Images** obtained using different design choices in the proposed workflow for five flights over the Cynthia cutblock. The best values are emboldened.

Design Choice	Jul 20	Jul 26	Aug 09	Aug 17	Aug 30	Mean
Unregistered	0.0539	0.0473	0.0695	0.0589	0.0658	0.0591
Manual	0.2417	0.1781	0.2612	0.2247	0.2003	0.2212
ECC	0.1658	**0.2141**	0.1742	0.2051	0.2079	0.1934
NGF	**0.2999**	0.2003	**0.3181**	**0.2715**	**0.3038**	**0.2787**
Perspective	0.2994	0.1995	0.3176	0.2707	**0.3052**	0.2785
Affine	**0.2999**	**0.2003**	**0.3181**	**0.2715**	0.3038	**0.2787**
Single-resolution	0.0239	0.0323	0.0271	0.0262	0.0247	0.0268
Multi-resolution	**0.2999**	**0.2003**	**0.3181**	**0.2715**	**0.3038**	**0.2787**
Batch size = 1	0.0420	0.0396	0.0602	0.0506	0.0477	0.0480
Batch size = 4	0.1420	0.0759	0.1053	0.1343	0.1036	0.1122
Batch size = 16	**0.3003**	0.1966	0.3176	0.2672	0.3017	0.2767
Batch size = 32	0.2966	0.1982	**0.3182**	0.2679	0.2999	0.2762
Batch size = 64	0.2999	**0.2003**	0.3181	**0.2715**	**0.3038**	**0.2787**
Random sampling	**0.3000**	0.1963	0.3177	0.2682	0.3023	0.2769
Systematic sampling	0.2999	**0.2003**	**0.3181**	**0.2715**	**0.3038**	**0.2787**

2.3.2 Qualitative Results

The quality of the generated thermal orthomosaic depends on the accuracy of the geometric alignment between the individual RGB and thermal images. If the co-registration is poor, some tree crowns appear twice while others are missing from the orthomosaic. Additionally, objects tend not to line up correctly. Fig. 2.8a shows the RGB orthomosaic generated for the Cynthia cutblock from the August 30 flight. Fig. 2.8b shows the orthomosaic obtained by texturing using unregistered thermal images. Although there are no gaps (similar to the RGB orthomosaic), the poor quality of this thermal orthomosaic is noticeable from the jagged nature of the vertical paths

through the trees (cutlines). Using the co-registered images obtained after applying the transformation matrix computed with NGF and the other best-performing design choices denoted in Table 2.2, a higher quality orthomosaic is generated, as shown in Fig. 2.8c. The cutlines are straight, and no individual trees are missing or duplicated.



(a) RGB Orthomosaic

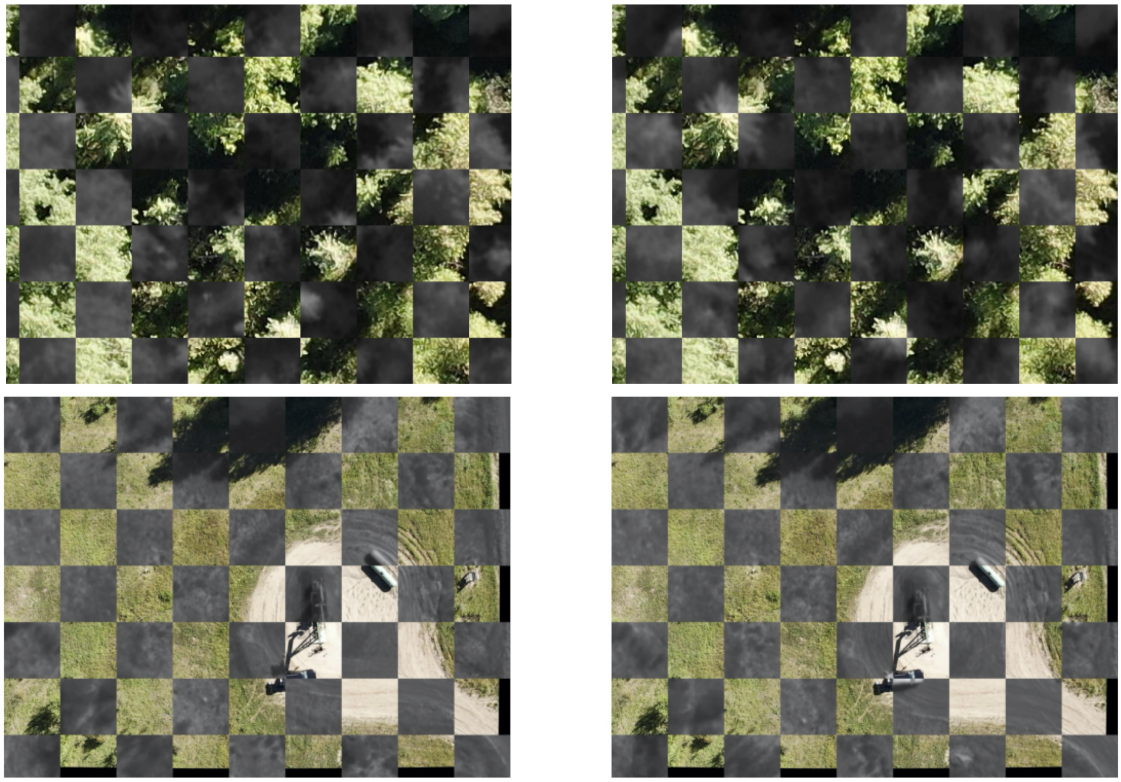
(b) Thermal Orthomosaic from Unregistered Images

(c) Thermal Orthomosaic from NGF-registered Images

Figure 2.8: **Visualization of Orthomosaics** generated from August 30 Cynthia cutblock data. (a) Georeferenced RGB orthomosaic. (b) Georeferenced thermal orthomosaic from unregistered images. (c) Georeferenced thermal orthomosaic from NGF-registered images. The improved quality of the orthomosaic in (c) is especially evident from the circular inset showing a straight path between the trees (similar to (a)) compared to the jagged path in (b).

The co-registration performance can also be observed by interlacing an undistorted RGB image with its corresponding thermal image before and after co-registration, as shown in Fig. 2.9. Before co-registration, there is an offset between the images that is especially noticeable from the larger individual trees in the top row and the road and parked vehicles in the bottom row. The geometric alignment of the two images significantly improves after co-registration. This explains why the co-registration stage results in a higher quality orthomosaic when the images are used to texture the surface

mesh reconstruction. Because of the geometric alignment of the individual RGB and thermal images, the resulting orthomosaics are also geometrically aligned. Thus, the co-registration stage is crucial to the success of the proposed workflow.



(a) Images before registration

(b) Images after registration

Figure 2.9: **Checkerboard Visualization** of two samples of undistorted RGB images interlaced with their corresponding thermal images (a) before and (b) after performing image co-registration with the proposed workflow. Coloured squares correspond to the RGB images, and grayscale ones to the thermal images.

2.3.3 Robustness of Transformation Matrix Computation

As mentioned previously, five transformation matrices were computed for performing co-registration, one for each of the five flight dates. Since the same H20T instrument was used across all flights, the computed transformation matrices are expected to be identical in theory. In practice, variations may arise due to differences in lighting conditions during data acquisition, for example. Table 2.3 reports statistics of the average, minimum and maximum values of each component of the affine transformation

matrices computed for all five flights. The maximum coefficient of variation (CoV) was observed to be 33.69% for $M_{2,1}$ (the component that controls vertical shear). Despite this high relative value, the maximum absolute difference from the average for this component across the five flights was merely 0.00536, which has a minor impact on the final transformation. For instance, with the point (400,300) in the original image (approximately the middle point between the image centre and top left corner), applying the 3×3 transformation matrix corresponding to the highest value of $M_{2,1}$ yields the point (390.36, 269.04). On the other hand, applying the matrix with the smallest value of $M_{2,1}$ yields (391.74 266.85). The Euclidean distance between these 2D image points is just 2.59 pixels or around 13 cm on the ground. This confirms that the co-registration used in the proposed workflow is robust, with only minor differences in the computed transformation matrices across different flights.

Table 2.3: **Robustness of NGF-based Co-registration**. Average, minimum, and maximum observed values over all flights for each affine transformation M component are listed. The coefficient of variation (CoV) for each component is reported in the final column. $M_{i,j}$ denotes the value in the ith row and jth column. $M_{3,1}$ and $M_{3,1}$ are always 0 for affine transformation matrices and thus omitted.

Component	**Average**	**Minimum**	**Maximum**	**CoV (%)**
$M_{1,1}$	1.01442	1.01380	1.015200	0.05
$M_{1,2}$	0.00618	0.00579	0.006600	4.45
$M_{1,3}$	0.02537	0.02075	0.030240	12.19
$M_{2,1}$	-0.00861	-0.01397	-0.005990	33.69
$M_{2,2}$	0.94546	0.94442	0.946730	0.08
$M_{2,3}$	-0.06670	-0.08146	-0.049870	16.40
$M_{3,3}$	1.04259	1.04153	1.043970	0.09

2.3.4 Radiometric Analysis

Here, I show that the proposed workflow preserves the radiometric information in the individual thermal images used to generate the orthomosaic. Specifically, I show that the captured absolute temperature values are the same before and after orthomosaicking for any region. Fig. 2.10 shows the central 256 × 204-pixel region of a sample thermal image and its corresponding patch (of the same size) extracted from the thermal orthomosaic generated by the proposed workflow. Cropping was done for this experiment to ensure that the orthomosaic patch corresponds to only this single image, as the texture mapping tends to only use the central portion of images during orthomosaic generation. The figure also shows that their temperature histograms (in bins of $0.1°C$) are nearly identical. The level of similarity between the histograms can be quantified using the Bhattacharyya coefficient [8], which measures overlap between two statistical populations. For 50 randomly chosen thermal images across all five flights, the average Bhattacharyya coefficient with their corresponding orthomosaic patches was 0.992 (minimum 0.984, maximum 0.998). This is very close to the theoretical maximum value of 1. Therefore, the orthomosaicking workflow successfully preserves radiometric information, i.e., absolute temperature values.

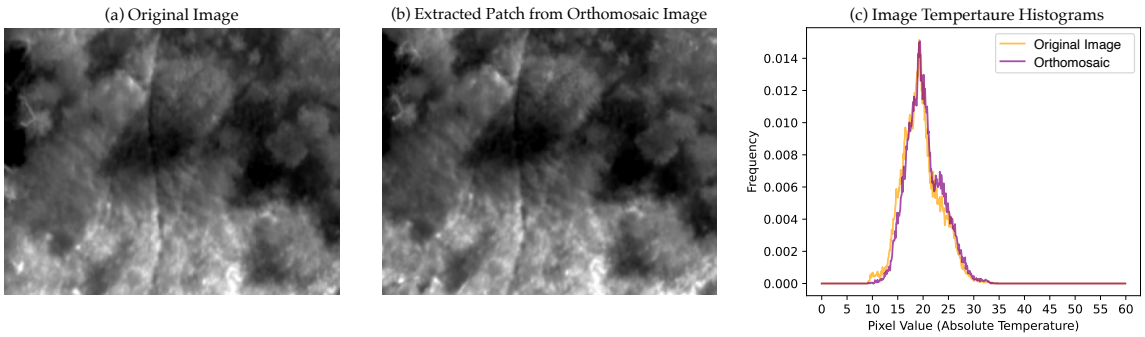
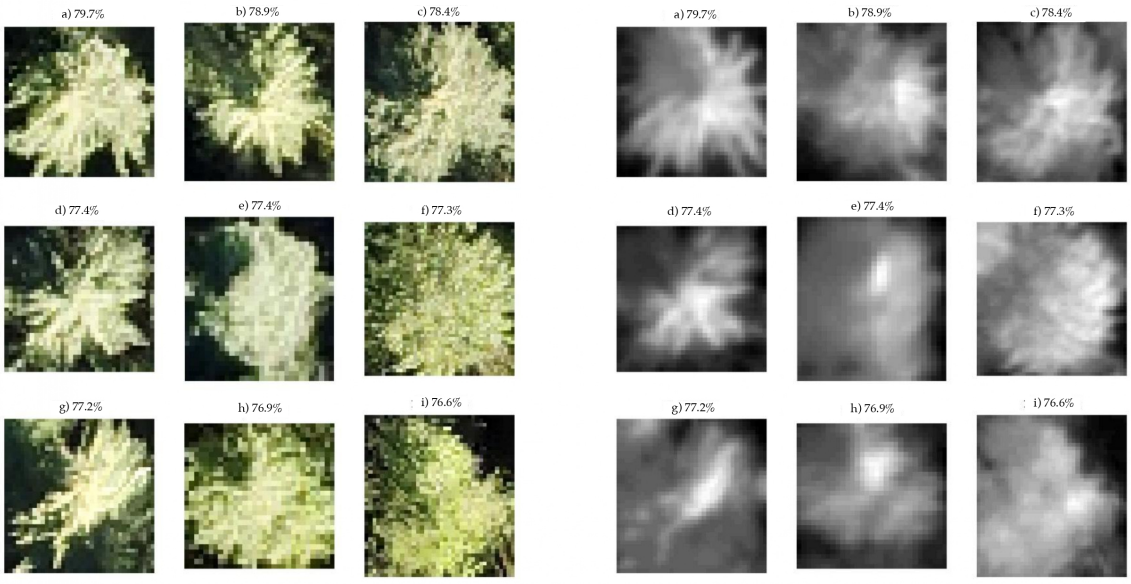


Figure 2.10: **Visual and Radiometric Similarity** between thermal images and the thermal orthomosaic. (a) A thermal image, (b) its corresponding patch in the thermal orthomosaic generated from the proposed workflow, and (c) temperature histograms for both in equally-spaced bins of $0.1°C$.

2.3.5 Downstream Task Performance

Fig. 2.11a shows some examples of the detected tree crown patches from RGB orthomosaics using the DeepForest pre-trained tree crown detector. The confidence score of each prediction in the figure was greater than 75% – there is only one tree centred tightly within each patch. In total, there were 8729 trees detected with confidence over 50%. The same bounding box coordinates were used on the thermal orthomosaic, and the extracted patches are shown in Fig. 2.11b. As a result of the good geometric alignment between the generated orthomosaics, the same trees appear in each of the corresponding pairs at the same locations. Hence, the proposed workflow enables applying an external model trained solely on RGB images to correctly detect individual tree crowns from the generated thermal orthomosaic, without having to train the detector on this new modality.



(a) Tree crowns detected from the RGB orthomosaic.

(b) Patches extracted using the same pixel locations as in (a) from the thermal orthomosaic.

Figure 2.11: **Visualization of Downstream ITCD Task Performance on Orthomosaics** generated from the August 30 Cynthia data. (a) Detected tree crowns from the RGB orthomosaic. The percentages represent the detection model's confidence score for each RGB crown. (b) Corresponding patches extracted from the thermal orthomosaic at the same pixel locations as in (a).

2.3.6 Processing Time

The CPU-only processing time for 814 RGB images (cropped to 1622×1216) and 814 thermal images (640×512) for the August 30 flight through all stages of the proposed workflow was around 2 hours and 25 minutes (with around 75 minutes for co-registration). This was for a machine running on a 12th Generation Intel Core i9-12900K 3.19 GHz processor with 32GB of RAM. With a single NVIDIA GeForce RTX 3090 GPU, the total time was reduced to around 80 minutes. The bulk of this processing time was taken up by the ODM process with RGB images for obtaining the surface mesh and undistorted RGB images. The co-registration stage took just around 10 minutes, meaning that the thermal orthomosaic can be generated in a fraction of the time taken for the RGB orthomosaic due to the reuse of intermediate outputs.

2.4 Discussion

The quantitative results reported in Section 2.3.1 emphasize the importance of choosing the correct hyperparameters during the RGB-thermal image co-registration stage. Automated intensity-based co-registration through gradient descent of the NGF loss function on average outperformed the other co-registration techniques, owing to its suitability for gradient descent optimization and multi-modal data [34]. Using ECC as the loss function resulted in a higher MI for one of the flights (July 26) but performed poorly compared to NGF overall. This indicates that ECC as the loss function in the proposed workflow is not as robust to different data as NGF for co-registering RGB and thermal drone images of forests. Manually supplying point correspondences for co-registration performed slightly better than ECC but worse than NGF, possibly due to human errors in selecting the exact pixels for correspondence. These errors stem from the possibility that hotspots (i.e., bright points) in a thermal image may not correspond to easily identifiable features in the RGB counterpart and vice versa. Regarding the transformation matrix, restricting it to six degrees of freedom (affine) yielded a slightly

better mean MI than allowing all eight degrees of freedom (perspective). This is because the RGB and thermal sensors lie on the same plane (or at least parallel planes) in the drone instrument, capturing nadir images. Hence, there is only a variation in the scale and translation, with possible rotation and skew between the two sets of images [63]. Allowing the two additional degrees of freedom in perspective transformation matrices led to a non-zero level of warping in the z-direction (perpendicular to the image plane). While this did not significantly reduce MI (and actually increased MI for the August 30 flight), it was observed to introduce a regular distortion pattern in the generated orthomosaic due to the misalignment of the image planes. Hence, the proposed workflow only considers affine transformation matrices.

The quantitative results further show that using a multi-resolution Gaussian pyramid framework is essential for proper co-registration - only relying on the single-scale original image significantly deteriorated performance compared to even the unregistered images. This is consistent with previous research showing that multi-resolution frameworks benefit image registration [78]. It can also be seen from the results that batch processing consistently outperformed single-image processing due to the computation of average gradients during optimization that reduces variance and promotes good convergence [106]. Using larger batch sizes increased performance for all flights, and although a batch size of 16 or 32 yielded competitive results with a size of 64, the latter option performed the best on average across all flights. Finally, the results corroborate the intuition that systematic sampling of image pairs for batch processing offers a good starting point for successfully performing co-registration. Compared to random selection, systematic selection was more robust, resulting in a slightly higher mean MI value. Overall, the reported results justify the specific design choices of the intensity-based co-registration stage within the proposed orthomosaicking workflow by demonstrating the robustness and high performance of the automated co-registration using the NGF loss function, multi-resolution image pyramid framework, batch processing, and systematic sampling. The highest MI of 0.2787 was achieved

during my experiments using this co-registration framework. This may seem low compared to the average MI of 1.4896 for the blue and red channels of the undistorted RGB images. However, this can be explained by the fact that thermal and RGB images carry very different information, so the MI for two perfectly co-registered optical images (i.e., red and blue channels of the same RGB image) has to be much higher than that of a precisely co-registered optical-thermal pair.

The qualitative results shown in Section 2.3.2 confirm that the selected design choices for the co-registration workflow, which individually were the best-performing options in terms of MI, collectively yielded a high-quality thermal orthomosaic that is geometrically aligned with its RGB counterpart. The co-registration of individual RGB and thermal images facilitates the correct reuse of intermediate outputs from the RGB orthomosaicking process to bypass the more problematic initial stages of the thermal orthomosaic generation. As a result, gaps and swirling artifacts [138] were absent in the generated thermal orthomosaic, demonstrating that the proposed integrated workflow can overcome these common issues of thermal-only processing workflows. The orthomosaic generated from unregistered thermal images was of poorer quality because the lack of geometric alignment of the individual RGB-thermal pairs renders the camera orientations inapplicable to the unregistered thermal images during texturing. Despite variations in conditions such as lighting across different flights, the five transformation matrices computed for RGB-thermal co-registration showed only minor variances as reported in Section 2.3.3, thereby additionally demonstrating that the workflow is robust across different flight data. The results in Section 2.3.4 further show that the thermal orthomosaic generation preserves radiometric information - absolute temperature values were unchanged by the orthomosaicking process. This is evidenced by the intensity histograms of image patches before and after orthomosaicking appearing highly similar and yielding a Bhattacharyya coefficient close to 1, the theoretical maximum. Additionally, the ITCD performed in Section 2.3.5 confirms that the generated orthomosaics are geometrically

aligned while simultaneously demonstrating the value of this alignment. The perfect match of the RGB crown boxes when applied to the thermal orthomosaic arises from the geometric alignment performed during the RGB-thermal image co-registration stage of the proposed integrated workflow. If the individual images are unregistered, the generated thermal orthomosaic has inconsistencies, such as the jaggedness of straight cutlines. Another less-noticeable yet significant issue of improper co-registration was that some trees in the study area became missing while others were duplicated.

In the remainder of this section, I discuss additional recommendations to effectively utilize the proposed workflow (and developed tool). Within the workflow, it is required to first crop the central regions of the RGB images. This can be done by specifying an appropriate scale in the GUI tool. For the H20T camera considered in this work, the 1622×1216 central region was cropped out of the original 4056×3040 wide-angle RGB images (i.e., 40% of width and height). When working with wide-angle RGB images, this has the added benefit of preventing the barrel distortion near the edges of such images from propagating to the generated orthomosaic (so that there is no visible tree lean) without significantly reducing the overlap between successive images. This results in a high-quality RGB orthomosaic, as shown in Fig. 2.8a.

The implementation of the thermal extraction stage within the GUI tool is specific to thermal images captured using a camera supported by the DJI Thermal SDK (e.g., Zenmuse H20 series, Matrice 30 series, and DJI Mavic 3 enterprise). When using the tool with these cameras, the input thermal images can be the unprocessed RJPEG files. There is an option for unsupported cameras to skip this stage and instead directly specify the path to the converted thermal images that should be used. These images should be similar to the 32-bit floating TIFFs output by the DJI Thermal SDK, i.e., images representing temperature values.

In a multi-resolution framework, the exact number of levels and the downscale factor also affect co-registration performance. A good minimum size for the width of the smallest image level is 20 pixels, so for a given downscale factor d and original

image width w, the number of levels L should be set to $\lceil \log_d(\frac{w}{20}) \rceil$. Based on this formula, I found that $d = 1.5$ and $L = 11$ performed well for the Cynthia cutblock data. Additionally, the success of co-registration depends on the learning rate and number of gradient descent iterations during optimization. These typically vary depending on the dataset, but I found 200 iterations at a fixed learning rate of 0.005 to be sufficient to reach convergence in all experiments, using the Adam optimizer [59]. Besides co-registration performance, hardware memory constraints and processing time are important considerations. Datasets with more images or with larger image dimensions necessitate longer processing times. Larger image dimensions also require more hardware memory, and so a smaller batch size may be needed – in my experiments a batch size of 32 or even 16 led to competitive results and should be safe alternatives.

An important choice I made in this work is that a single linear transformation matrix was used to co-register all thermal and undistorted RGB image pairs for a given flight data. The results showed that this choice yielded good thermal orthomosaicking performance in terms of both quality and geometric-alignment with the RGB orthomosaic. An alternative, pair-specific diffeomorphic co-registration can be done, as in [34] for medical images. This non-linear warping has the advantage of accounting for any uncorrected differential distortions present in the image. However, it has significant drawbacks that prevent its effective application for orthomosaicking. First, it is not robust for a given flight - since it would require computing a specific transformation for each pair, if any of the registrations performed worse than the others, that part of the orthomosaic would be of poorer quality and possibly even unusable. Second, it is not computationally efficient to compute a non-linear transformation for each pair, especially for longer flights having more image pairs.

2.5 Conclusions and Future Work

This chapter proposed a new workflow that generates two geometrically-aligned orthomosaics from simultaneously acquired RGB and thermal drone images. Compared

to previous workflows that process thermal data separately and hence generate lower-quality orthomosaics suffering from gaps and swirling artifacts, the proposed workflow leverages the intermediate outputs of RGB orthomosaic generation and only uses thermal images for texturing, thereby overcoming those issues. Using an automated intensity-based image co-registration, the method achieves good geometric alignment between the individual thermal and RGB images, which allows for using the thermal images to properly texture the surface mesh previously reconstructed from the RGB images. The co-registration optimizes the NGF loss function that is based on image gradients and is found to outperform ECC, an alternative registration technique commonly used in previous works. The co-registration of the individual images translates to the geometric alignment of the two generated orthomosaics. This is advantageous for downstream forest monitoring tasks, as demonstrated by the tree crown bounding boxes detected from the RGB orthomosaic by a DL model being directly applicable to the same tree crowns in the thermal orthomosaic generated from the proposed workflow. Moreover, the orthomosaicking process preserves the radiometric information present in the original thermal images. To facilitate future research and improvement of the proposed workflow, a flexible open-source tool with an easy-to-use GUI has been developed and is publicly available[1] to facilitate use by practitioners.

The generated orthomosaics in this chapter provide a comprehensive representation of the AOI and can be used for a variety of downstream analysis and processing tasks, as exemplified by the ITCD performed briefly in Section 2.3.5. The next chapter details a novel ITCD model that outperforms the RGB-only detector used in this chapter, especially under challenging illumination conditions.

[1] https://github.com/rudrakshkapil/Integrated-RGB-Thermal-orthomosaicing

Chapter 3

Shadow-Agnostic Tree Crown Detection

3.1 Introduction

Forest environments are important to ecosystems, economies, and society worldwide. A critical step in forest remote sensing is individual tree crown detection (ITCD), which can assist ecologists, foresters, biologists, and land managers in increasing the scope of their sampling for performing tasks such as pest infestation detection [54, 83], carbon storage estimation [26], and species identification [5, 96], and various other forest health monitoring applications [21]. In recent years, various DL-based ITCD methods have been proposed to address the challenges in forest monitoring [151]. However, the lack of diverse, publicly available datasets tailored to this specific application has impeded progress in this research domain. Additionally, the ITCD task poses significant application-oriented and environmental challenges. These challenges include effectively harnessing the information from multiple sensors and ensuring the robustness of results in the presence of environmental factors. Existing tree crown detectors (e.g., [136]) have primarily been trained on RGB images, which are sensitive to occlusions and illumination variations (e.g., for shorter trees hidden in shadows). Nevertheless, the advantages of incorporating thermal images with complementary information in ITCD have been largely overlooked. While a few studies have used RGB-thermal data for urban tree crown detection (e.g., [89]), they

require extensive manual pixel-wise annotation for supervised training and fail to address forest monitoring challenges, such as shadowed or occluded tree crowns. To bridge these gaps, this chapter aims to provide an aligned RGB-thermal forest tree crown dataset. It proposes a novel self-supervised approach that leverages both RGB and thermal imagery, improving the accuracy and adaptability of ITCD in various illumination conditions.

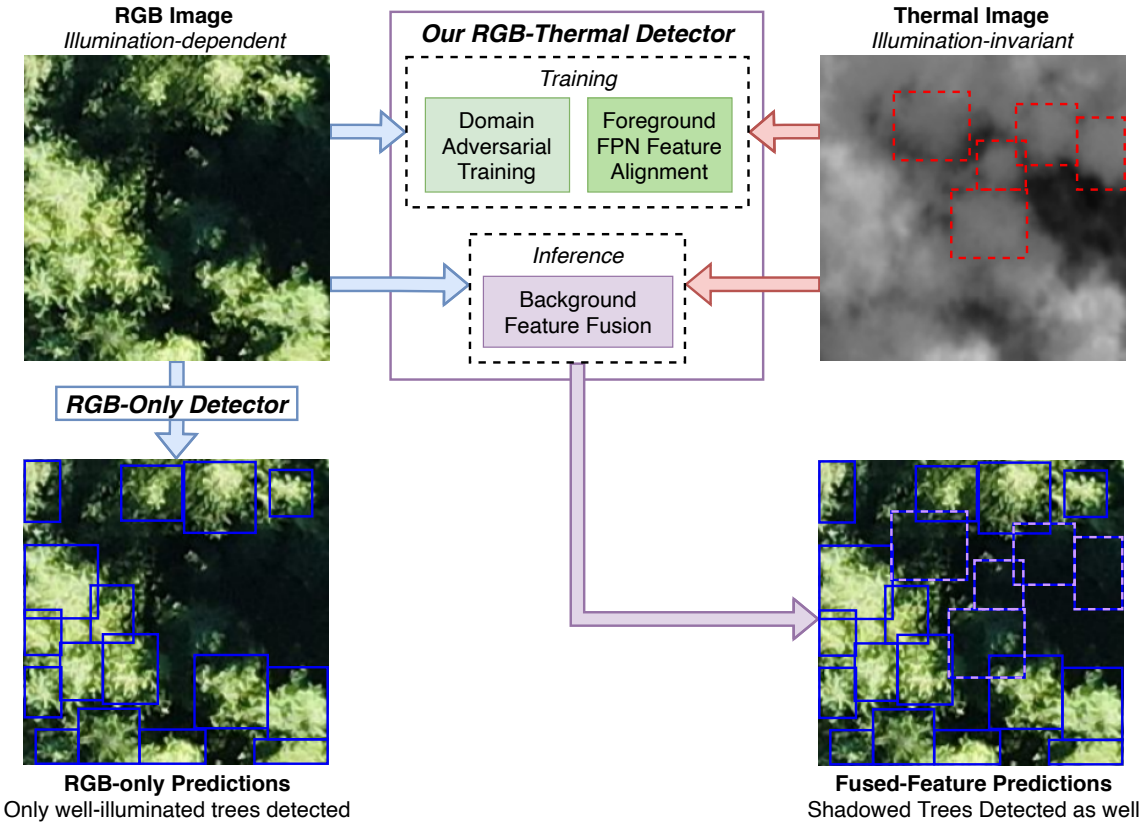


Figure 3.1: **Overview of Proposed Method**. Undetected trees hidden in shadows are indicated by dotted red boxes. Best viewed in color.

The proposed method, named ShadowSense, comprises domain adversarial training (DAT) and foreground (FG) feature alignment to learn domain-invariant representations and match observed tree crowns in both modalities (see Fig. 3.1). In particular, I train a shadow-agnostic ITCD model consisting of two parallel branches based on the RetinaNet architecture [67]. After initializing both branches with RGB-trained weights of the detector, the thermal branch and three domain discriminators are jointly trained,

minimizing the domain discrimination loss for the feature extractor while maximizing it for domain discriminators. Tree crowns visible in both modalities are adapted by aligning FG feature maps of the feature pyramid network (FPN) using a simple yet effective intensity-based segmentation followed by morphological operations. During inference, the background (BG) regions of the FPN feature maps from RGB-thermal modalities are fused using a weighted average. The fused maps are then passed to the detector heads, leading to accurate prediction of tree crown bounding boxes. The proposed method is entirely self-supervised, avoiding the need for labour-intensive manual annotations for model training. Moreover, a challenging, large-scale dataset consisting of registered RGB-thermal drone image pairs is presented, serving as a valuable resource to develop robust models and support future research.

My main contributions in this chapter are summarized as follows,

1. I propose a novel shadow-agnostic tree crown detection method to exploit complementary information of RGB-thermal images and overcome the limitations of recent RGB-trained models for the ITCD task. The method leverages the registered nature of available data for self-supervision (i.e., eliminating the need for data annotations) and incorporates source domain data post-adaptation.

2. I provide a challenging dataset for shadowed tree crown detection encompassing varying degrees of shadows and illumination conditions in a complex forest environment. This RGB-thermal dataset is large-scale and includes annotated images for evaluation, and unlabelled images for training, aiming to advance the development of unsupervised/self-supervised methods.

3. I perform extensive empirical evaluations to demonstrate the superior effectiveness of the proposed method when compared to SOTA methods that utilize image-to-image translation, early image fusion, or unsupervised domain adaptation (UDA).

3.2 Related Work

Tree Crown Detection. DL methods have gained significant popularity in ITCD from RGB drone imagery in recent years. These methods primarily rely on well-known object detectors with different architectures [39, 151] and have found applications in various domains (e.g., [84]). However, these models are often trained on small datasets, resulting in moderate performance and the inability to effectively address challenges like overlapping canopies, small tree crowns, and distractors in various forest environments. Among the existing ITCD methods, DeepForest [136] stands out as the SOTA detector and was trained on a manually annotated dataset comprising over 10k tree crowns from 37 forests across the United States of America. Nevertheless, despite its remarkable performance in well-illuminated conditions, this RGB-only trained detector struggles to accurately detect trees with crowns hidden in shadows.

Unsupervised Domain Adaptation (UDA). The goal of UDA is to transfer knowledge from a source domain (e.g., RGB) to a target domain (e.g., thermal) without relying on annotations specific to the target domain [99]. In general, UDA involves adapting models either within the same modality (e.g., RGB with clear vs. foggy weather) or across different modalities, such as RGB-thermal [3, 27, 58, 86, 111, 129, 153].

Many UDA approaches incorporate DAT by integrating domain discriminator networks into multiple parts of the model to encourage learning domain-invariant representations. These methods often employ global domain classification for the entire image [100], or local pixel-wise classification focusing on FG regions [145] or areas of interest predicted by attention modules [58, 129]. Despite their success for generic object detection/segmentation, the potential application of these methods for the ITCD task is still largely unexplored. A drawback of applying existing UDA methods to this task is their reliance on source domain annotations for training (see Table 3.1), which is not feasible in this problem setting. To address this limitation,

the method I propose in this chapter adapts to the thermal data distribution without requiring any annotations, utilizing only the registered nature of the available data for self-supervision. Moreover, unlike previous approaches, my method retains and incorporates the source domain data after adaptation.

RGB-Thermal Early Fusion. Instead of adapting an RGB-trained model to the thermal data, an alternative approach is to fuse information from both modalities into a more informative image. Most works cope with the lack of ground truth fusion results by employing unsupervised RGB-thermal fusion methods. These methods can be applied to unregistered or registered images. For instance, UMFusion [132] improves upon existing methods for unregistered images by incorporating style transfer and a parallel-branch fusion module. Wang *et al.*[135] propose an attention-based method to integrate thermal target perception and RGB detail characterization for scenarios with registered images (like in this work). These methods perform fusion at the image level, resulting in a new, richer image with combined properties from both modalities.

Alternatively, fusion can be conducted at the intermediate feature level. Moradi *et al.*[89] propose a U-Net-based fusion model for tree crown segmentation, which requires ground truth segmentation maps for training. Supervised intermediate feature fusion methods have also been extensively studied in tasks like classification [62], segmentation [64, 121], and salient object detection [32, 65, 120, 127, 133, 148, 155] (SOD, see Table 3.1). In contrast, the proposed method performs feature fusion during inference, rather than early fusion at the image level, in a self-supervised manner specifically designed for ITCD.

Image-to-Image Translation. Aside from fusion approaches, an alternative research direction involves colourizing a thermal image to resemble its RGB counterpart using encoder-decoder networks [16, 47, 76], and leveraging the 'translated' image for downstream tasks. Another approach is the use of classical algorithms [25] or SOTA DL methods [31] to translate RGB images into shadow-free versions. However, these

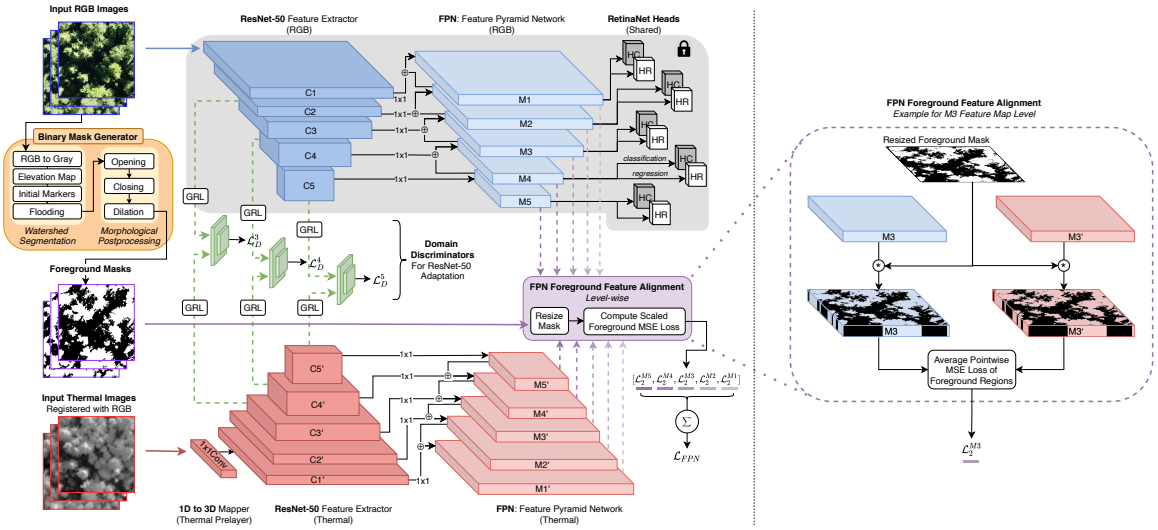
methods typically discard the original RGB images in preference of the translated images, which may suffer from image artifacts and potentially contain less semantic information. Instead, the proposed method effectively fuses intermediate features extracted from both modalities after performing the UDA, thereby preserving the complementary information of RGB and thermal modalities.

Table 3.1: **Categorization of Related Works** according to training supervision through RGB ground truth (GT) annotations.

Category	GT Required	GT Not Required
UDA for Detection	[56] [57] [77] [86] [90] [100] [130] [129] [145] [153] [154]	**Proposed**
RGB - Thermal Fusion	[29] [69] [70] [89] [152] *SOD*: [32] [65] [120] [127] [133] [148] [155]	[38] [68] [132] [135] [141] and **Proposed**
Translation	[16] [31]	[47] [76]

3.3 Proposed Method and Dataset

In this section, "visible trees" refers to trees seen in both RGB and thermal images, primarily due to adequate lighting conditions. In addition, "shadowed trees" are commonly shorter trees that remain hidden by the shadows of neighbouring larger trees in the RGB image but become apparent in the thermal image. Due to the limitations of the illumination-dependent RGB modality, RGB-trained detectors are ineffective in identifying a significant number of shadowed trees. This is primarily because signals beyond the visible spectrum are imperceptible using RGB sensors alone. Hence, the proposed method first adapts the backbone of the existing baseline detector to the thermal data and then fuses extracted features from both modalities during inference. In the remainder of this section, I present the proposed method in detail and then introduce an RGB-thermal dataset that facilitates advancements in challenging illumination conditions and enables the development of robust models for the ITCD task.



(a) Complete Training Workflow

(b) FPN Foreground Feature Alignment

Figure 3.2: **(a) Detailed Workflow of Proposed Training Procedure** consisting of a thermal branch (in red) and an RGB branch (in blue). The weights of both are initialized from [136], and the RGB branch is frozen during training. The thermal feature extractor is trained to fool the domain discriminators (in green), and vice versa, using gradient reversal layers (GRL) at multiple levels. **(b) Close-up** of FPN feature alignment (in purple) at the M3 level that encourages foreground feature map regions of the two branches for a given image pair to match.

3.3.1 Model Architecture and Training

The proposed model consists of two parallel branches (i.e., RGB and thermal branches) based on the RetinaNet architecture [67] for the detection task (see Fig. 3.2a). This architecture consists of a backbone network and detection heads. The backbone includes a ResNet-50 network [43] that extracts features at multiple resolutions from input images and an FPN that combines extracted features from multiple levels. Each branch of the proposed model comprises the backbone network, while the classification and regression heads that produce detection outputs are shared. A 1×1 convolutional layer (i.e., 'pre-layer') is attached at the beginning of the thermal branch for expanding thermal input images to three channels before passing them to the backbone network. Both branches are initialized with weights from a pre-trained RGB tree crown detector [136], and the RGB is frozen to maintain its performance in the source domain. This

is done because these weights effectively identify tree crowns from RGB images but perform poorly for thermal images, as indicated in Table 3.3. Then, DAT and FG FPN feature alignment are employed to adapt the thermal branch to the target domain distribution. Considering the thermal data's inherent low texture and contrast, the proposed training helps the thermal branch provide accurate predictions for visible and shadowed tree crowns.

Domain Adversarial Training (DAT). DAT is employed to train the thermal ResNet-50 feature extractor and thermal pre-layer to learn domain-invariant representations. Inspired by [100], three domain discriminator networks (shown in green in Fig. 3.2) are attached to the 3rd, 4th, and 5th levels of the extractor. These convolutional neural network (CNN)-based classifiers predict the domain label (i.e., RGB or thermal) for the computed feature map for input images during training. Each discriminator is preceded by a gradient reversal layer (GRL) [28] that acts as the identity function in the forward pass, i.e., $G(\mathbf{x}) = \mathbf{x}$, but negates gradients in the backward pass. This layer ensures that the gradients flowing through the extractor and the discriminators are in opposition. Doing so sets the stage for a two-player game: the feature extractor is trained to generate indistinguishable feature representations by discriminators between the source and target domains, while the discriminators aim to accurately classify the domain labels based on the feature representations.

The single-class focal loss is used to emphasize challenging images during DAT, as,

$$\mathcal{L}_D^c = -(1 - p_t)^\gamma log(p_t), \tag{3.1}$$

where c $\in \{3, 4, 5\}$ is the level of the feature map, p_t is the predicted domain probability, and γ controls the diminishing rate of the modulating factor. A larger weight is assigned to more challenging instances, increasing their importance in the overall loss calculation. Then, the game is modelled as a min-max optimization,

$$\min_{\{\theta_d^3, \theta_d^4, \theta_d^5\}} \max_{\{\theta_r, \theta_p\}} \mathcal{L}_D^3 + \mathcal{L}_D^4 + \mathcal{L}_D^5, \tag{3.2}$$

where $\theta_d^c, c \in \{3, 4, 5\}$ are the parameters of the three domain discriminators, θ_r are the parameters of the thermal feature extractor, and θ_p are the parameters of the thermal pre-layer. Unlike typical UDA works [90, 100, 129], the proposed method does not combine adversarial loss with a task-aware detection loss due to the lack of source annotations.

Foreground FPN Feature Alignment. It is crucial for FPN outputs of the RGB and thermal branches to align (i.e., be the same) for the trees that are visible in both modalities (i.e., in the FG regions). This alignment acts as a proxy for task-specific detection loss to guide adaptation during training and is vital for the weighted average fusion process during inference (see Section 3.3.2). Thus, the proposed method can ensure the effective combination of complementary information from both modalities, leading to improved detection performance for shadowed tree crowns. Fig. 3.2b illustrates the alignment process for the third FPN feature map. To do so, a binary BG/FG mask (obtained as described below) is down-sampled to the feature map size at the current level and then applied to the feature maps from the two branches. Standard average pixel-wise $L2$ loss is then computed between the residual values. Accordingly, five loss values denoted as $L_2^f, f \in \{1, 2, 3, 4, 5\}$ are obtained. These losses are then combined in a scaled manner, with higher weightage assigned to the larger feature maps using scaling values $\beta^f, f \in \{1, 2, 3, 4, 5\}$, i.e.,

$$\mathcal{L}_{FPN} = \sum_{f=1}^{5} \beta^f \mathcal{L}_2^f \, . \tag{3.3}$$

This alignment is complementary to the UDA process – both have the effect of producing the same feature maps at FG regions regardless of the modality. Therefore, $\mathcal{L}_{FPN}$ is used to update the parameters θ_f of the thermal FPN as well as the preceding parameters θ_r and θ_p.

To generate the binary masks used to train the detection model, a simple yet computationally efficient method combining classic watershed segmentation [124] and mathematical morphology is employed. This approach avoids the complexity of recent

methods that utilize auxiliary neural networks for mask prediction. It is particularly suitable for the shadow-agnostic ITCD task because it leverages the assumption that BG pixels (including shadows) are generally darker than those in the FG. According to the binary mask generator in Fig. 3.2, each RGB image is converted to grayscale. Then, pixels with intensity $< \frac{20}{255}$ are marked as 1 (representing the darker BG) and those with intensity $> \frac{100}{255}$ as 2 (i.e., brighter FG). The FG/BG labels for pixels with intensities in between (initially unmarked) are determined through Meyer's iterative flooding algorithm [7], as implemented in scikit-image [119, 131]. In this algorithm, an elevation map is computed using Sobel filtering. This map is then 'flooded' starting from the defined FG/BG markers. For this, each marked pixel's neighbours are inserted into a priority queue based on gradient magnitude, with enqueue time serving as a tiebreaker favouring the closer marker. The pixel with the highest priority is extracted, and if its already-marked neighbours share the same marker, it is assigned to that pixel. All unmarked neighbours that are not yet in the priority queue are enqueued. This flooding procedure iterates until the queue is empty and all pixels are marked as either FG or BG. After obtaining the initial binary mask, three morphological operations are applied for further refinement. Specifically, 4-connected 3×3 structuring elements are used for (1) *opening* to remove errant FG pixels surrounded by BG, (2) *closing* to remove errant BG pixels surrounded by FG, and (3) *dilation* to pad FG boundaries and maintain FG performance during inference.

3.3.2 Feature Fusion during Inference

During the inference phase, complementary information from the thermal branch is exploited to address the limitation of detecting shadowed tree crowns with only the RGB branch, thereby improving the overall ITCD performance. This information resides in the BG regions of the RGB images, which are typically prominent in their thermal counterparts. To achieve this, the same binary mask generation process is used as in the training phase, but this time assigning '1's to represent BG regions

and '0's for FG. Subsequently, the feature maps extracted from the RGB and thermal modalities are fused level-wise. In Fig. 3.3, I illustrate this fusion process for the M2 level of feature maps.

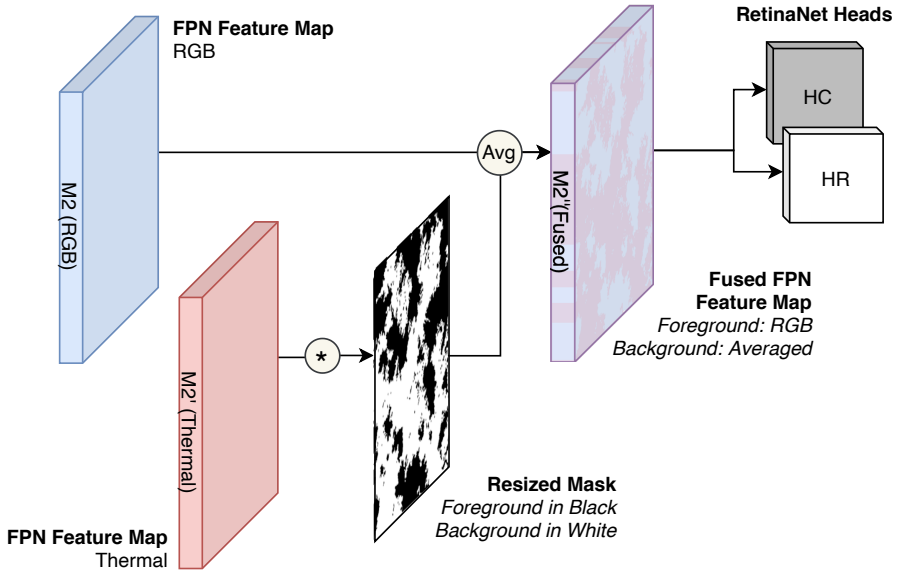


Figure 3.3: **Masked Fusion During Inference** for the M2 level feature maps as an example. Background features (purple) are obtained by weighted averaging of the RGB (blue) and thermal (red) features. Foreground features are assigned the original RGB values. Best viewed in colour.

While the FG pixels (depicted as black regions in Fig. 3.3) from the RGB feature maps are directly utilized, the BG regions of the thermal feature maps are masked to focus solely on the areas that are not visible in the RGB modality. As a result, the fused feature map F^f_{Fused} at level f is obtained through a weighted average of the RGB feature map (F_{RGB}) and the thermal feature map (F_T) for all BG pixels (x, y) as,

$$F^f_{Fused}(x,y) = \frac{F_{RGB}(x,y) + (F_T(x,y) \times \lambda_T \times \eta^f)}{1 + (\lambda_T \times \eta^f)}, \tag{3.4}$$

where λ_T is the weight assigned to thermal features for all levels and η^f denotes the fusion weight scaling specific to that level. η^f decreases with f because larger feature maps have a higher spatial resolution, and thus the averaging is less error-prone due to containing more fine-grained information. Once the fused feature map is obtained

at each level, it replaces the RGB feature map and is fed into the classification and regression heads to predict bounding boxes.

3.3.3 Dataset for Shadowed Tree Crown Detection

In this section, I present an RGB-thermal dataset titled *RT-Trees* for advancing shadowed tree crown detection and developing robust models for ITCD. This dataset builds on the existing data introduced in Section 2.2 by conducting additional flights using the same setup. Specifically, the same DJI H20T sensor was employed to capture RGB-thermal drone imagery during nine flights over a mixed forested region of central Canada. This data was then combined with available data from the original five flights. During data collection, flight times were purposely diversified to encompass a spectrum of challenging illumination conditions. Additionally, ever-changing climatic conditions throughout the year (e.g., temperature and snow cover) introduce an additional layer of diversity and challenges, especially in the more sensitive thermal images.

A series of preprocessing steps were applied on the raw drone imagery from all flights, including cropping, resizing, co-registration with NGF-based workflow (proposed in previous chapter), splitting into training/validation sets based on GPS coordinates, and providing high-quality annotations for evaluation purposes. This resulted in a substantial collection of approximately 50k registered image pairs across all flights, signifying a considerable expansion compared to existing RGB-thermal datasets (see Table 3.2). 63 non-overlapping images were sampled for testing and 10 for validation from a single flight date (August 30). Each tree crown only appears once in these sets to ensure the reliability of performance evaluation, and the annotations differentiate between visible and shadowed (i.e., "difficult") tree crowns. The remaining bulk of images (49,806) was designated for training. These images display a high degree of overlap ($> 75\%$) and span all flights, a deliberate choice aimed at promoting diversity and consequently justifying the discrepancy in data split numbers. RT-Trees is primarily intended for self-supervised RGB-thermal ITCD, so no training

set annotations are provided, but the proposed method demonstrates that the co-registered imagery can facilitate domain adaptation and feature fusion techniques through the registered nature of the multi-modal imagery. A notable characteristic of the RT-Trees dataset is the highly dense spatial distribution of detection targets compared to existing datasets, averaging around 60 tree crowns per image. Moreover, the presence of different tree species results in considerable variability of crown areas and shapes. The challenges of RT-Trees are substantiated with descriptive statistics in Appendix A, where I also describe the collection, pre-processing, and annotation procedures in more detail.

Table 3.2: **Comparative Overview** of RGB-Thermal Image Datasets.

Dataset	**# Pairs**	**Dimensions**	**Year**	**GT**	**Application**
TNO [125]	63	Various	2014	×	Image Fusion
MFNet [33]	1569	640×480	2017	✓	Semantic Segmentation
VIFB [149]	21	Various	2020	×	Image Fusion
RoadScene [142]	221	768×576	2020	×	Image Fusion
LLVIP [50]	15488	1080×720	2021	✓	Pedestrian Detection
M^3FD [70]	4200	1024×768	2022	✓	Object Detection
RT-Trees (Proposed)	52869	500×500	2023	✓(eval.)	Tree Crown Detection

3.4 Experiments

In this section, I first provide implementation details for the proposed method (ShadowSense). I then compare its performance with the baseline and existing SOTA methods through the quantitative results reported in Table 3.3. Specifically, I utilized three metrics for evaluation: (1) AP50, representing the average precision at 50% IoU (Intersection over Union) threshold, (2) AR100, representing the average recall over several IoUs given 100 detections, and (3) Percentage of correctly identified shadowed trees. The third metric focuses only on the difficult boxes, counting a positive if a predicted box with an overlap of 85% with the BG regions was assigned to a difficult box. Finally, I present qualitative comparisons.

3.4.1 Implementation Details

The well-known RGB-trained crown detector DeepForest [136] was considered the baseline method. The RetinaNet networks [67] in each RGB/thermal branch were initialized with pre-trained weights from [136]. To ensure fair comparisons, the RetinaNet hyperparameters were configured similarly to those employed in the baseline. This involved setting the non-maximum suppression threshold to 0.15 and the score threshold to 0.1 (default in [136]). During training, I set the FPN alignment scales $\beta = [1.0, 1.0, 0.5, 0.05, 0.01]$ and the focal loss parameter γ to 2 (recommended in [67]). The domain discriminators consisted of three *Conv-BatchNorm-ReLU-Dropout* layers, followed by an adaptive average pooling layer to reduce feature maps to a single channel and a linear layer to finally produce a single output representing the confidence of belonging to the target domain. Dropout layers with a probability of 0.5 were included for regularization. To suppress noisy classification signals during early training stages, the adaptation factor for the GRL was gradually increased from 0 to 1, as recommended in [28]. A training batch size of 16 was used in all experiments. The Adam optimizer [59] was used with an initial learning rate of 0.001, which was exponentially decayed with a gamma factor of 0.9 after each epoch (i.e., a complete pass through the training set). The training was conducted for 10,000 iterations, a sufficient period to observe plateauing in all training losses. The experiments were conducted on a single Nvidia GeForce RTX 3090 GPU with 24 GB of RAM. During inference, weighted fusion was performed using a thermal weight of $\lambda_T = 5$, which provided the best results. Similar to β, the scaling weights $\eta = [1.0, 1.0, 0.5, 0.2, 0.2]$ were applied to weight more towards thermal features in larger feature maps while also ensuring that all products of λ_T and η are greater than or equal to one (i.e., always at least equal weighting between thermal and RGB features). Further validation of selected hyperparameters is provided in Appendix B.

3.4.2 Baseline Quantitative Comparison

I evaluated the performance of the baseline model [136] in four scenarios. The first two involved assessing the effectiveness of this model off-the-shelf on RGB images and thermal images, respectively. The performance on RGB images was 49.86% AP50 and 24.01% AR100, although just 10.41% of the difficult shadowed trees were successfully identified. On thermal images, the baseline detector exhibited significantly inferior performance (see Table 3.3). These results demonstrate that the off-the-shelf RGB-trained detector is ill-suited for the thermal domain. In the other two scenarios, I conducted supervised fine-tuning of the detector model on RGB imagery, using supervised focal loss [67] for 10 epochs, following [136]. I used a subset of RT-Trees comprising 326 non-overlapping RGB images containing over 22.5k crowns of visible and shadowed trees. These were manually annotated for this experiment by inspecting the RGB-thermal image pairs. The performance on RGB images shows a lead of 5.34% and 5.41% in terms of AP50 and AP100, respectively, while also resulting in a 9.69% increase in the detection of shadowed trees. Although the thermal modality is not directly used for training, this configuration requires costly annotation based on both modalities. Also, the performance of this model on thermal images is dramatically poor due to low spatial resolution and lack of fine details in these images. Instead, the proposed ShadowSense can achieve superior performance by leveraging multi-modal data without needing *any* annotations during training at all.

3.4.3 State-of-the-art Quantitative Comparison

I compare the performance of the proposed ShadowSense with various image-to-image translation, RGB-thermal early fusion, or UDA SOTA methods. The baseline detector was applied to the generated images in the image translation and fusion methods. In contrast, the proposed weighted-average fusion of BG FPN feature maps was adopted in all UDA experiments (and ShadowSense) for fair comparisons. Additionally, I present an ablation study to analyze the impact of different components on ShadowSense.

Table 3.3: **Quantitative Comparison** of the proposed method with baseline and SOTA methods using % AP50 (↑), % AR100 (↑), and % of shadowed trees correctly identified (↑). Best and second-best results are emboldened in red (supervised) and blue (self-supervised).

Evaluation		Method	Training Data	All Trees		Shadowed Trees
				AP50	AR100	% Identified (↑)
Baseline [136]		Off-The-Shelf Model (Eval. on RGB Images)	NEON [137] (RGB only)	49.86	24.01	10.41
		Off-The-Shelf Model (Eval. on Thermal Images)		4.34	2.27	2.82
		Supervised Fine-Tuned Model (Eval. on RGB Images)	+ Ann. RGB RT-Trees subset	**55.20**	**29.42**	**20.10**
		Supervised Fine-Tuned Model (Eval. on Thermal Images)		5.64	3.98	3.81
Image Translation Inference using [136] on Generated Images		Increased Background Brightness in HSV Space	N/A	42.01	20.35	10.41
		ShadowFormer [31]: Shadow Removal	ISTD [134]	11.62	6.48	3.47
		PearlGAN [76]: Thermal Image Colourization	RT-Trees	40.72	19.67	11.18
Early Image Fusion Inference using [136] on Generated Images		UMFusion [132]	TNO [125]	38.00	18.56	10.20
		MFEIF [68]	TNO [125]	39.62	18.95	15.40
		MetaFusion [152]	M^3FD [70]	43.17	21.29	18.00
RGB-Thermal UDA without Source Domain Annotations Our Fused Inference after Adaptation		SSTN [90]: Contrastive Learning	RT-Trees	31.53	15.16	2.13
		Attention-based UDA [129]	RT-Trees	31.97	15.34	3.84
		DA-RetinaNet [100]: ResNet DAT	RT-Trees	32.88	15.72	3.65
	Ablation Study	(i) **Proposed**: FG FPN FA	RT-Trees	47.11	22.48	5.21
		(ii) **Proposed**: ResNet DAT + FPN FA w/o Masking	RT-Trees	49.75	23.22	10.63
		(iii) **Proposed**: ResNet DAT + FG FPN FA (Pred. Masks)	RT-Trees	52.18	24.84	9.33
		(iv) **Proposed**: ResNet FG DAT + FG FPN FA	RT-Trees	52.24	24.38	14.32
		(v) **Proposed** (**ShadowSense**): ResNet DAT + FG FPN FA	RT-Trees	**54.13**	**25.76**	**19.09**

Image-to-Image Translation. I investigated the effectiveness of three such methods: PearlGAN [76] (SOTA thermal colourization method); ShadowFormer [31] (SOTA shadow removal method); and a classic method that increases the brightness of pixels in HSV colour space proportionally to their original brightness, i.e., darker pixels are made brighter. Thermal images colourized using PearlGAN performed worse than the baseline by -9.14% AP50 and -4.34% AR100. However, slightly more shadowed trees were detected. The decreased performance can be attributed to the introduction of artifacts and an overall loss of semantic information compared to the original RGB images. The classical shadow removal method showed better results than PearlGAN but performed worse than the baseline. This method jitters the entire image inconsistently with the detector, leading to poor performance. The detection performance on images generated by ShadowFormer was the most inadequate, achieving an AP50 of 11.62% and an AR100 of 6.48%, with just 3.47% shadowed trees identified. This model is ineffective for removing the shadows of dense tree canopies in the RT-Trees dataset.

RGB-Thermal Early Fusion. I evaluated three SOTA methods: UMFusion [132], MFEIF [68], and supervised MetaFusion [151]. MetaFusion directly generates a fused three-channel RGB image, whereas UMFusion and MFEIF convert the RGB image to the YCbCr colour space, fuse the brightness (Y) channel with the thermal image, and then convert the fused image back to the RGB space. In all three methods, shadowed trees become partially visible in the fused images to varying extents. According to the results, MetaFusion achieved the best performance. Although the AP50 and AR100 results were still lower than those of the baseline detector, the fused images revealed 7.59% more shadowed trees than the baseline. Similar trends were observed for UMFusion and MFEIF. Although RGB-thermal early fusion improved the visibility of BG regions, the performance of detecting FG tree crowns deteriorated. In summary, the overall detection performance of all three methods was worse than the baseline.

UDA. As shown in Table 3.1, existing UDA methods require ground truth annotations to compute task-specific detection loss during training, which guides the adaptation process. To ensure a fair comparison, three UDA methods compatible with the one-stage object detector RetinaNet were selected – Attention-based UDA [129], SSTN [90], and DA-RetinaNet [100]. These methods were modified by excluding only the supervised detection loss due to the lack of training annotations in RT-Trees. In Attention-based UDA, the attention module that dynamically selects local feature regions for adaptation was trained using DAT alongside the thermal branch of the proposed model. In the case of SSTN, only the ResNet-50 and pre-layer of the thermal branch were fine-tuned using contrastive loss as described in [90]. Similarly, for DA-RetinaNet [100], only global DAT (i.e., for all regions of extracted feature maps) was employed to adapt the ResNet and pre-layer of the thermal branch. Among these three methods, DA-RetinaNet demonstrated the best adaptation to the thermal data distribution. However, its performance was still limited as numerous false positive predictions contributed to the overall insufficient performance. The drawback of these methods lies in the absence of task-aware detection loss during adaptation due to the

lack of ground truth annotations in RT-Trees. Consequently, these models cannot learn to extract domain-invariant representations that are meaningful for the detection task. Instead, they primarily learn to deceive the domain discriminators irrespective of the downstream task (i.e., detection).

The proposed method overcomes the limitations of the discussed UDA methods by incorporating task-aware FG FPN feature alignment (abbreviated FG FPN FA) to guide the adversarial adaptation process. By using DAT and FG FPN FA, the proposed method outperforms the baseline RGB-only detector and existing SOTA methods with an AP50 of 54.13% and AR100 of 25.76%, with 19.09% of shadowed trees successfully detected (almost doubling the success rate of the baseline). The entirely self-supervised method performs comparably to the supervised fine-tuning method without requiring labour-intensive manual labelling. The proposed feature fusion process selectively enhances features in the BG regions using thermal-extracted features. Importantly, this fusion does not have an adverse effect on FG performance, which distinguishes ShadowSense from existing early fusion approaches. Additionally, the fusion process leverages available data from both domains for detection, unlike single-domain image-to-image translation methods.

3.4.4 Ablation Study

A systematic ablation analysis of the proposed method is presented in Table 3.3. It includes five different configurations: (i) FG FPN FA using the proposed classic image masking (CIM) with no DAT applied to the ResNet model, (ii) ResNet DAT with FPN FA and no masking (aligning all regions of feature maps), (iii) ResNet DAT with FG FPN FA and different masking (using baseline detector predictions as FG and the rest as BG), (iv) Pixel-wise ResNet DAT (discriminators output domain labels for each pixel and consider loss only for FG pixels) with FG FPN FA using CIM, and (v) Proposed ResNet DAT with FG FPN FA using CIM (ShadowSense).

According to the results, the following key inferences can be made: 1) UDA

through DAT is crucial (from (i) & (v)): relying solely on FG FPN FA led to inferior performance compared to the baseline, indicating that the thermal branch did not effectively learn to extract domain-invariant representations without DAT. 2) FG masking for FPN alignment is crucial (from (ii), (iii) & (v)): Even with DAT, aligning whole feature maps slightly decreased performance compared to the baseline – aligning features of a tree visible only in the thermal image with BG features from the RGB image interfered with training. 3) The proposed mask generation method outperforms RGB detector-predicted mask generation (from (iii) & (v)): The proposed masking detected significantly more shadowed trees than this alternate masking, showing superior performance for FPN FA and fusion. 4) FG masking is unnecessary for DAT (from (iv) & (v)): Pixel-wise domain classifiers with loss computation restricted to FG regions resulted in slightly worse performance than global DAT, likely due to the usage of less available data (only FG pixels vs. all pixels) in the same number of training iterations.

3.4.5 Qualitative Results

Feature Space Visualization. Fig. 3.4 shows the initial disparity between the RGB-thermal FPN feature map representations before the proposed training procedure. After training, however, they become indistinguishable as domain-invariant feature maps are aligned and thus can be directly averaged for fusion during inference.

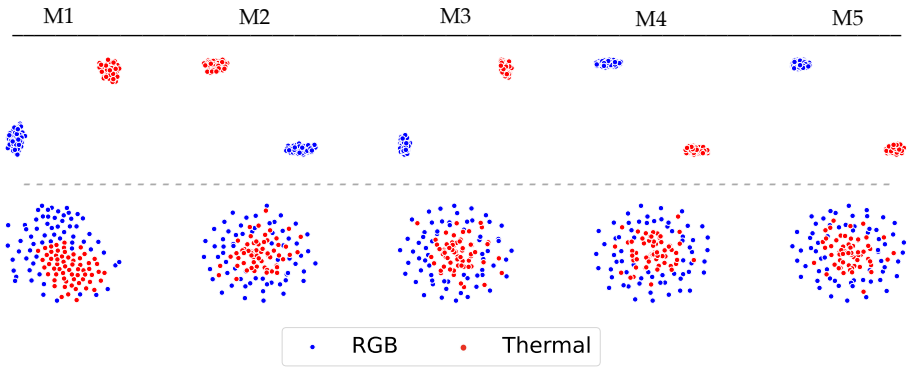


Figure 3.4: **t-SNE Visualization** of RGB-thermal FPN features: (top row) before training and (bottom row) after training.

Detection Performance. The visual performance of the proposed method is compared to the off-the-shelf RGB detector [136] for five different image patches in Fig. 3.5. Two outputs of the proposed method are shown: one from thermal FPN feature maps (isolated thermal branch) and the other from the fused feature maps. The thermal branch can detect shadowed trees in the BG that were missed by the baseline, but there is a decline in the FG performance. In the latter, the BG detections are accurately propagated while maintaining the FG baseline performance. Overall, the proposed method outperforms the baseline by comprehensively improving the detection results.

3.5 Conclusions

In this chapter, I proposed a novel shadow-agnostic ITCD method and presented a challenging RGB-thermal dataset to address the limitations of existing RGB-only detectors. The proposed method exploits DAT and FG FPN FA to learn domain-invariant representations and match visible tree crowns between RGB and thermal modalities. Unlike existing adaptation methods, the approach does not require annotations for task-aware supervision during training. Instead, it relies on the registered nature of the images for aligning feature maps of visible FG regions. The proposed method effectively detects shadowed trees by fusing complementary thermal information. Further, the dataset presented comprises registered RGB-thermal drone image pairs that can stimulate future research in challenging ITCD scenarios. Experimental comparisons demonstrate the superiority of the proposed method over the baseline RGB-trained detector as well as SOTA image fusion and UDA-based techniques.

Although the experiments in this chapter consider drone-collected images to ensure a larger training set, it is entirely possible to apply the proposed detection model to the registered orthomosaics generated in the previous chapter (recall that the baseline RGB-only detector was used in this chapter was the same as that in the previous one). The next chapter explores a downstream classification task on the tree crown patches. These patches can be effectively extracted using the proposed model in this chapter.

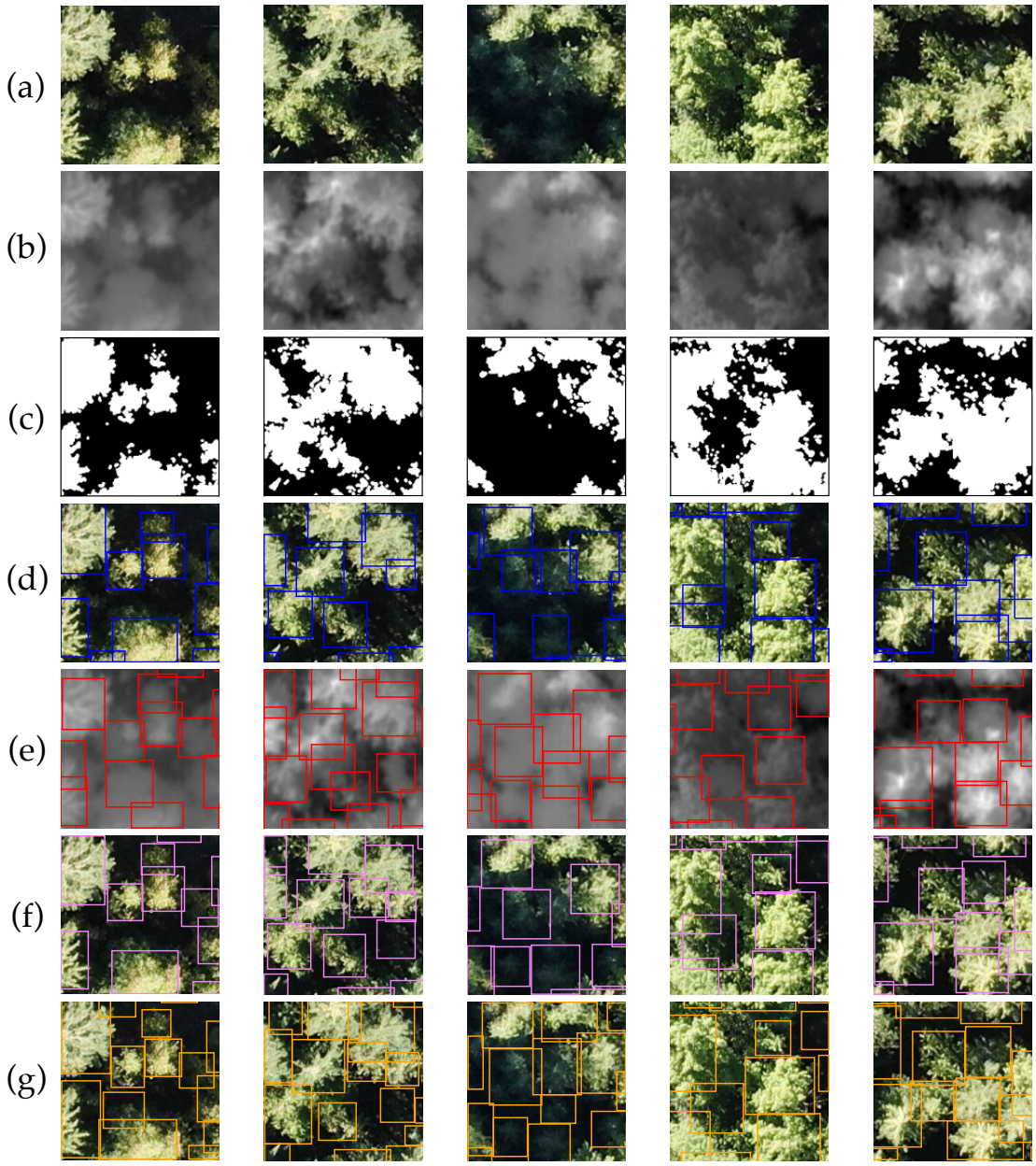


Figure 3.5: **Tree Crown Detection Results**. Each column shows (a) RGB image, (b) Thermal image, (c) Generated mask; and predictions by (d) Baseline [136], (e) DAT-adapted thermal branch, (f) Proposed ShadowSense, and (e) Ground truth. Best viewed in colour.

Chapter 4

Bark Beetle Visible Attack Stage Classification

4.1 Introduction

Bark beetle outbreaks significantly impact forests worldwide, thereby disrupting the functioning and properties of natural ecosystems. As a result of various factors (e.g., population density, tree moisture & condition, beetle & host tree species), a successful bark beetle attack gradually reveals itself by affecting various parts of the host tree [22]. Over time, the crown of an infested tree begins to fade – there is a gradual change in foliage colour from a healthy green to yellow, red, and finally a leafless (i.e., needle-less) grey. These are referred to as different attack stages.

The crown fading process is linked to the life cycle of bark beetles (see Fig. 4.1), in which pioneer female bark beetles bore tunnels (called oviposition galleries) in the phloem of host trees to lay their eggs, and the larvae hatch and excavate additional larval galleries to feed on phloem tissue. The rate of discoloration depends on the progress of bark-beetle-induced fungal infection that further interrupts nutrient and water flow through the phloem and xylem of host trees, as well as environmental conditions such as soil moisture content [94]. Typically, the change from green to red takes one year [10]. However, variations in the host tree's defensive response to minimize water losses can significantly delay the onset of visible discoloration [94]. Suppose colonization is successful and the host tree's defences are overwhelmed. In

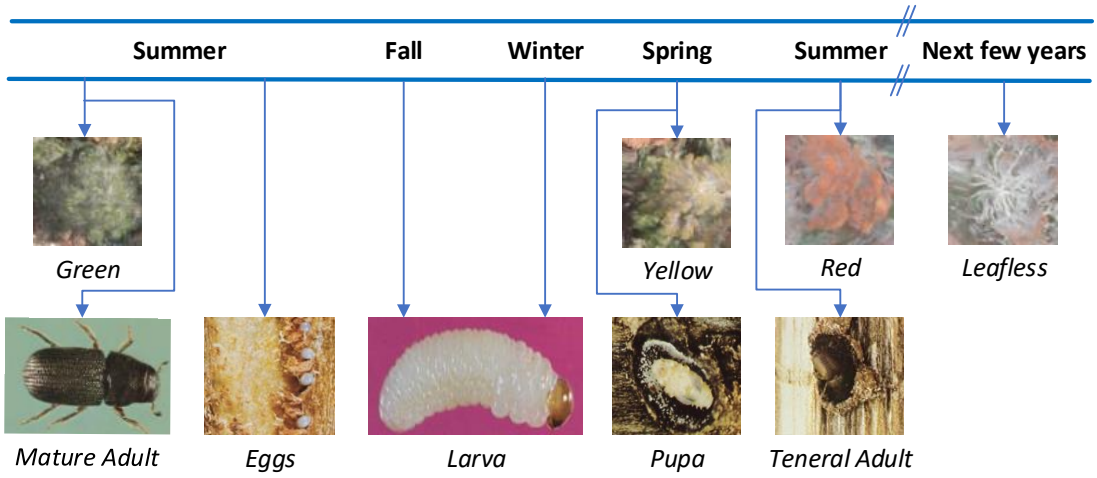


Figure 4.1: **Typical Life Cycle of Bark Beetles** and their effect on host tree foliage over time. Beetle images have been adapted from [109].

that case, the tree ultimately dies, and the next generation of beetles disperses from the parental tree in search of new hosts [109].

Emerging bark beetles from infested trees disperse in several ways in search of new hosts, with the majority partaking in short-range dispersal [110]. They fly below the forest canopy and attack suitable host trees within a few hundred meters. Hence, identifying previously colonized trees will help determine the next likely location of infestations and guide beetle management activities (e.g., sanitation, removal, or disposal) to prevent infestations from further spreading [37].

Bark beetle outbreaks have had devastating consequences [11]. Infestations affect the total volume of merchantable pine, even if some trees can be salvaged. Rural communities that rely on the forests for local employment and tourism are especially vulnerable. Beyond the risk to commercial interests, bark beetle-induced tree mortality can adversely affect priceless ecosystem service values, such as biodiversity and carbon sequestration. Increased wildfire risk is another significant impact since dead trees ignite more easily than live trees. These consequences highlight the importance of detecting bark beetle activity to mitigate their spread by falling and burning infested trees over the winter and early spring [13]. The detection of infested trees

by *Dendroctonus mexicanus* is studied in this chapter, which is among Mexico's most damaging insects for pine forests.

Traditionally, bark beetle attacks have been detected through satellite or aircraft platforms and classical ML-based approaches such as random forests (RF) or support vector machines (SVM). Although satellite and aircraft platforms are widely used at the landscape level, recent research has focused on leveraging UAVs for data collection due to their advantages at the individual tree level (e.g., higher spatial and temporal resolution). Besides, classical ML-based approaches require feature selection, which demands prior domain experience and extensive effort to achieve satisfactory results. Thus, exploiting DL-based models is of interest due to their capacity for learning powerful representations and exhibiting good generalization by discovering intricate, underlying data patterns. As shown in Fig. 4.2, an automated system that detects and analyzes bark beetle infestations using remote sensing and machine learning (ML) is desirable to avoid labour- and cost-intensive efforts of typically employed ocular assessments.

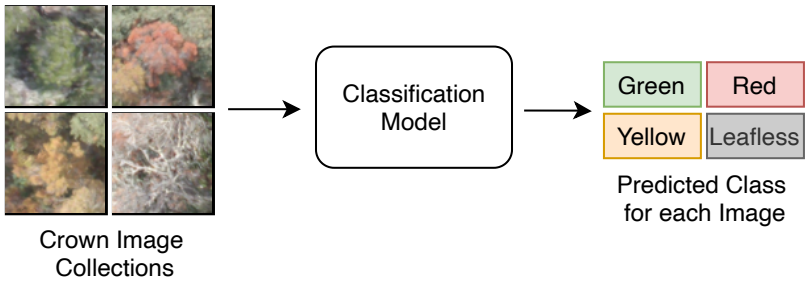


Figure 4.2: **Brief Diagram** of the desired classification model for this task.

4.2 Related Works

In this section, I briefly describe existing DL-based approaches that seek to detect trees infested by bark beetle species from UAV-captured images (i.e., individual tree level). First, the potential of deep neural networks (DNNs) to detect bark beetle outbreaks in fir forests was studied in [108]. A two-stage method consisting of a classical image

processing-based crown detection and a six-layer convolutional neural network (CNN) was employed for predicting red- and grey-attacked trees by four-eyed fir bark beetles (*Polygraphus proximus Blandford, Coleoptera, Curculionidae*). This method used RGB images captured by a DJI Phantom 3 Pro quadcopter, and the performance was compared to six well-known CNN models (e.g., VGGNet [116], ResNet [43], and DenseNet [46]). After that, the classification accuracy of infested trees in a temperate forest was investigated in [88] by training two shallow CNNs (with three and six convolutional layers) and applying transfer learning to a pre-trained DenseNet-169 [46]. Despite the availability of multi-spectral images from a DJI Matrice 210 RTK, the best results of this method were obtained using only RGB bands for detecting yellow-attacked trees. Finally, the health statuses of Maries fir trees were evaluated in [93] by adopting pre-trained CNN models of AlexNet [61], SqueezeNet [45], VGGNet, ResNet, and DenseNet. Using a DJI Mavic 2 Pro & DJI Phantom 4 Quadcopter, this method used RGB images to delineate treetops traditionally and classify healthy and grey-attacked trees.

In contrast, I propose to adapt a SOTA, deep RGB tree crown detector [67] (same as the RGB-only baseline detector used in the previous chapter) for the classification of bark beetle attacks by exploiting backbone network weights that have been specifically pre-trained for tree crown detection from UAV images, along with an introduced shallow subnetwork for distinguishing between attack stages.

4.3 Proposed Method

Even though this task seems like a simple colour classification, ill-defined attack labels and imbalanced datasets make it more challenging than it appears. For instance, the distribution and some challenging samples are visualized in Fig. 4.3, in which green and leafless (needle-less) classes overlap with other classes. Also, Fig. 4.4 shows the RGB colour space histograms for each class that reveals similarities between the yellow and red attack stages due to the gradual foliage discoloration.

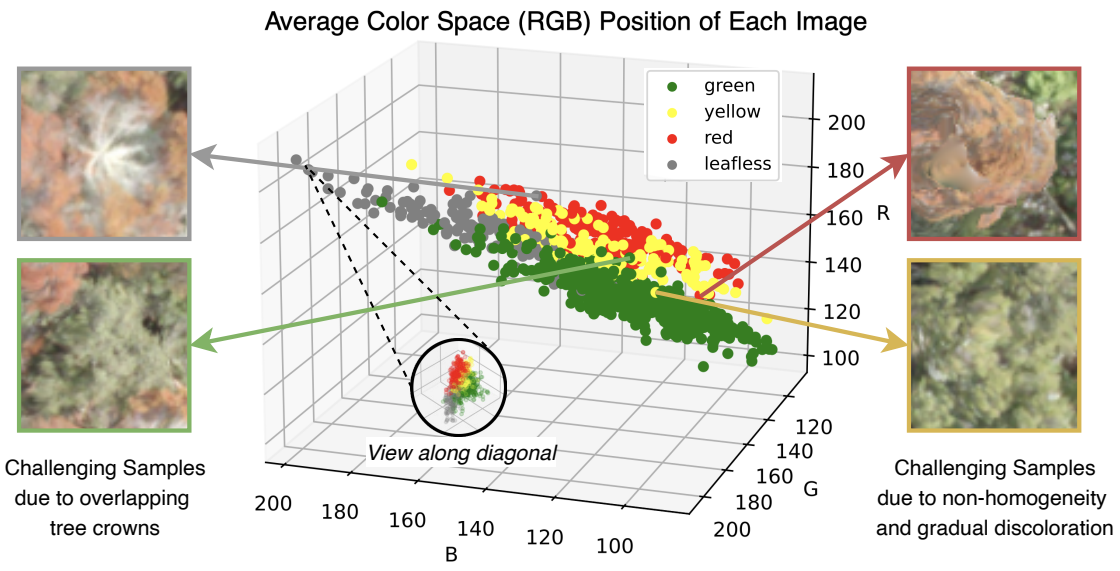


Figure 4.3: **RGB Colour Space Distribution of Bark Beetle Dataset Images**. The borders of the highlighted challenging samples indicate their true labels.

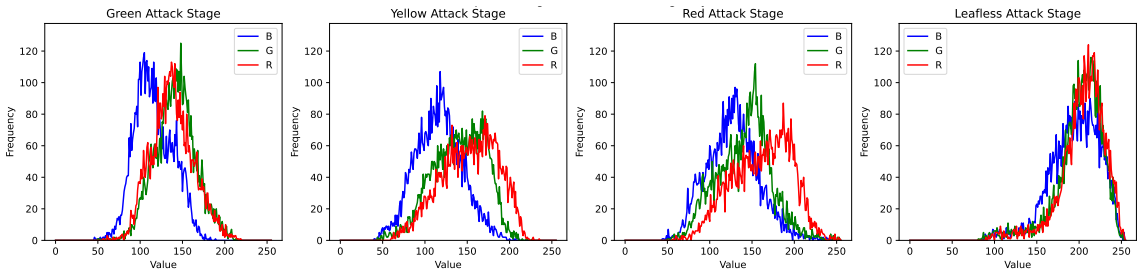


Figure 4.4: **Histograms** showing RGB colour space distribution of the different attack stages with leaves.

The proposed method is based on the RetinaNet architecture [67], which has been successfully used for other remote sensing applications (e.g., [146]) owing to its ability to detect dense targets from data with highly imbalanced classes. The proposed RetinaNet-based architecture includes a backbone network (i.e., ResNet-50 feature extractor [43] and feature pyramid network (FPN)), classification subnetwork, and focal loss. Although the backbone network seeks to extract multi-scale features, the FPN combines semantically low-resolution features with low-level, high-resolution ones. The classification subnetwork then predicts the category of bark beetle attacks (i.e., green (healthy) tree, yellow-/red-attacked tree, or leafless) using focal loss. This

loss function helps to simultaneously handle the inherent similarity of attack classes and limited data by focusing on hard samples and avoiding easy negatives.

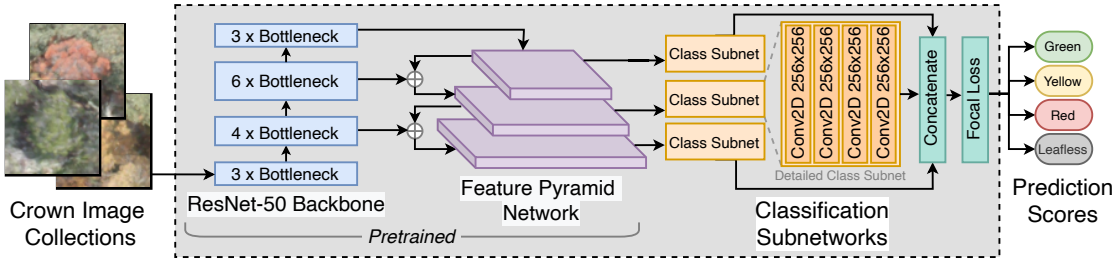


Figure 4.5: **An Overview of the Proposed Method**. First, the ResNet-50 and feature pyramid network (FPN) are initialized using the tree crown detection pre-trained baseline weights [136]. Following that, the network is modified and fine-tuned to classify the stages of bark beetle attack.

As shown in Fig. 4.5, the cropped images of tree crowns are normalized according to the mean and standard deviation of the training set images and then fed into the backbone network. The computations are forward propagated through the bottleneck layers and combined with the different layers in the FPN. Each level of the FPN feeds its computation to a classification subnetwork consisting of four convolutional layers. Then, the network outputs a score for each attack stage. At last, one-hot encoding is done to get the class prediction for each tree. Considering the available tree crown collections, the bounding box regression subnetwork typically found in the RetinaNet architecture has been removed. In contrast to previous studies that train either a shallow network or deep models pre-trained on ImageNet [18], the proposed method exploits a pre-trained deep model (i.e., DeepForest [136] for tree crown detection) and trains the modified network for the classification of attack stages. As a result of appropriately initializing the network with weights relevant to tree crowns, the classification subnetwork can focus on learning to differentiate between different bark beetle attack stages. To overcome class imbalance in the considered dataset and increase the total number of training images, data augmentation is employed prior to training. In this work, several different data augmentation strategies were considered. Although it is generally assumed that data augmentation will result in

better performance for DL models, the results in Section 4.4.3 show that blindly utilizing these techniques can drastically affect classification results for this challenging task.

4.4 Empirical Evaluations

I evaluated the proposed method using the dataset presented in [113], which utilized a hexacopter with a Tarot FY680 Pro to capture multiple RGB video sequences of a forested region in Northern Mexico from a top-down perspective. Five flights in total were conducted at three different average heights above ground (60m, 90m, and 100m) during three months (June, July, and August). The individual frames from each flight were combined into five different orthomosaics, and the ground truth information for each tree's centre and attack stage was available (see [113] for more details). The proposed method is compared with the baseline method [113] and the most promising SVM, RF, and K-nearest neighbours (KNN) classifiers. During the experiments, the hyperparameters for each classifier were tuned using grid search.

4.4.1 Implementation Details

As shown in Fig. 4.6, individual tree crowns were cropped from the five orthomosaics as patches of 76x76 pixels and split into five separate sets with training, validation, and testing subsets (see Table 4.1). One model was trained for each flight, and evaluation was performed for each individually and averaged. The proposed networks (five models for flights) were trained using the AdamW optimizer [74] for 50 epochs and a batch size of 2. The training procedure was performed on an Nvidia GeForce RTX 3090 GPU, with each model taking approximately 1.5 hours to train. The dataset was augmented by generating minority class samples using i) random affine warps, ii) vertical/horizontal flips, iii) 90°/180°/270° rotations, iv) cropping a random sub-patch of 70% and resizing, v) colour jittering with random brightness, contrast, & saturation, and vi) Gaussian blurring with kernel size 5, visualized in Fig. 4.7. Furthermore, early

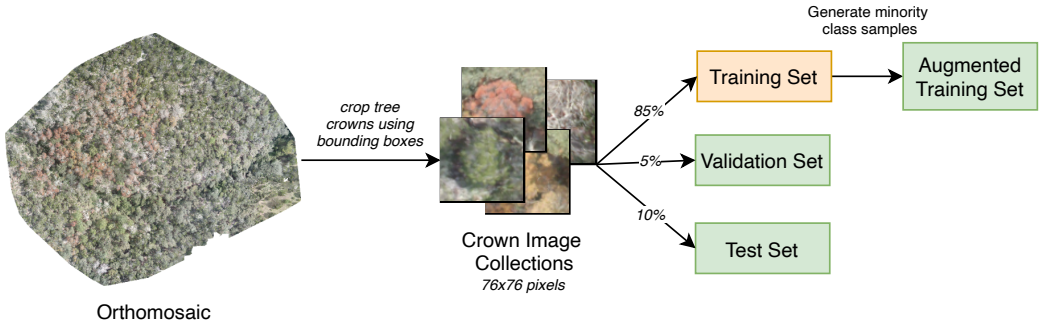
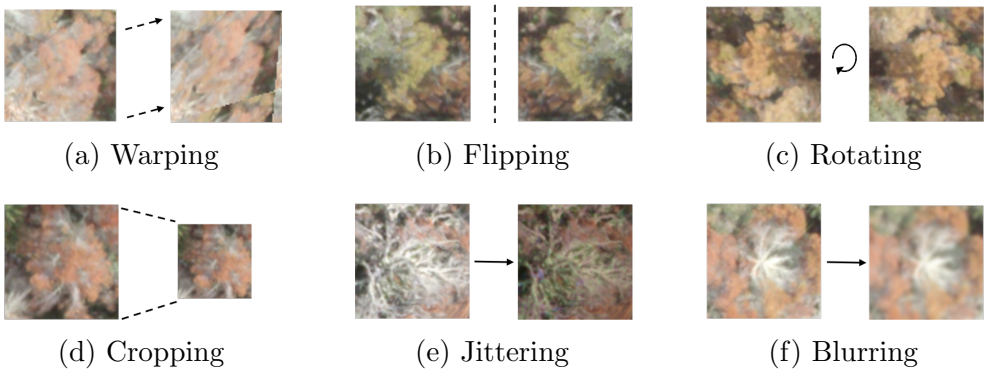


Figure 4.6: **Data Processing Pipeline** for the bark beetle attack stage classification orthomosaics.

stopping was considered to avoid overfitting during training.



(a) Warping (b) Flipping (c) Rotating

(d) Cropping (e) Jittering (f) Blurring

Figure 4.7: **Visualization of Data Augmentation Strategies** considered to produce minority class samples.

Table 4.1: **Dataset Distribution** for each flight according to attack stage label and training/validation/testing split.

Subsets of Samples	**Jun 60m**	**Jul 90m**	**Jul 100m**	**Aug 90m**	**Aug 100m**
Green Trees	68	81	103	141	98
Yellow Trees	34	19	28	45	49
Red Trees	24	26	48	52	48
Leafless Trees	25	28	26	33	25
Train	128	130	174	230	187
Augmented Train	232	276	352	480	332
Validation	7	7	10	13	11
Test	16	17	21	28	22

4.4.2 Experimental Results

The experimental comparison of the proposed method with the baseline and best-performing models for classical ML methods is shown in Table 4.2. The random guess accuracy for classification between four classes is 25%. According to the results, the proposed method with affine augmentation considerably outperformed the cellular automaton baseline method – by 9.9% (& 7.6%) with (& without) data augmentation in average accuracy across all flights. Also, the classical ML methods achieved significantly lower accuracy than the proposed method. This can be explained by the ill-defined separation between classes in the RGB colour space, as shown in Fig. 4.3. The confusion matrices for the challenging flights for the best-performing proposed model are shown in Fig. 4.8. There are no misclassifications for four of the flights, and only one leafless image is incorrectly predicted as red in the June 60m flight due to the considerable overlap from nearby red attack stage trees. Since classic ML methods rely on manual feature selection, applying them directly to raw data (like images of tree crowns) results in poor performance. However, the proposed DL-based method can automatically learn the most relevant and robust features from the dataset, enabling it to perform significantly better.

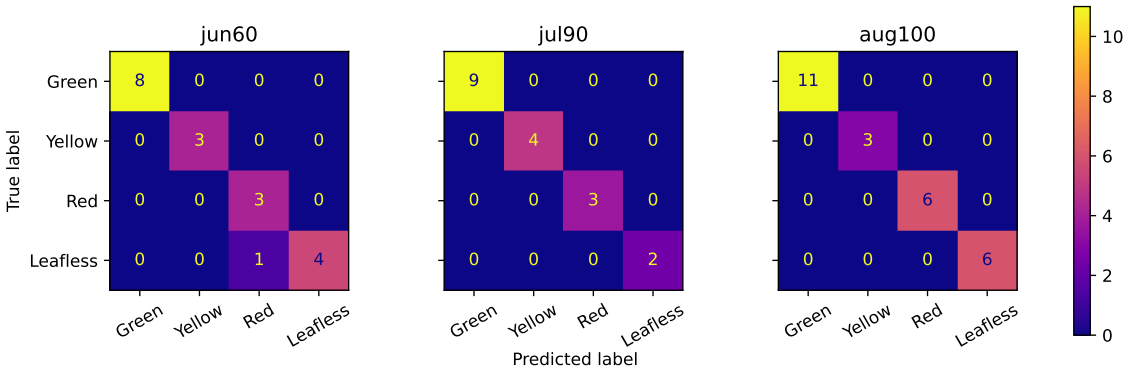


Figure 4.8: **Confusion Matrices** for the best-performing proposed model.

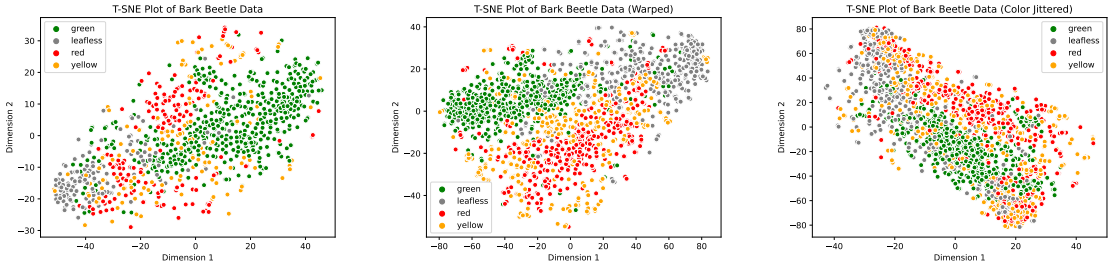
Table 4.2: **Classification Accuracy** for various attack stage classification models. The best result is emboldened.

	Model	Average Accuracy (↑)
	SVM	46.37%
	KNN	45.89%
	RF	35.49%
	Baseline [113] (Best result)	89%
	Proposed (with warping)	**98.95**%
Ablation Study	Proposed (without augmentation)	97.69%
	Proposed (with cropping)	96.29%
	Proposed (with flips)	94.74%
	Proposed (with rotation)	94.71%
	Proposed (with blurring)	92.23%
	Proposed (with color jittering)	83.90%

4.4.3 Ablation Study

Various probabilistic augmentation strategies were studied to assess the effectiveness of data augmentation. In each strategy, additional samples belonging to the red, yellow, and leafless classes were randomly generated to obtain the same number as the green samples and balance the dataset. The classification results for models trained on each strategy are shown in Table 4.2. Accordingly, affine warping was the most effective strategy considering tree crowns are not always circular. This strategy changes the apparent geometry of the trees, promoting more diversity in the dataset. Also, it accounts for angular variation in the UAV during data collection. Colour jittering unsurprisingly led to the most performance degradation, explained by its major effect on the images that further confuses the model between visual symptoms of trees. These results are further analyzed using t-SNE visualizations in Fig. 4.9. The middle and left plots display similar separations in the dataset, indicating that warping added

minority class samples without adversely impacting the separation of the classes. On the other hand, the right plot is obtained from the colour-jittered dataset, and significantly more overlap between the classes can be observed (e.g., in the bottom right corner). The other augmentation strategies did not improve performance either.



(a) No augmentation (b) Affine warping (c) Colour jittering

Figure 4.9: **t-SNE Visualization** of dataset with different augmentations.

4.5 Conclusion

A DL-based approach to classify bark beetle-infested trees is proposed in this chapter. Based on the RetinaNet architecture, the proposed method simultaneously trains a shallow subnetwork and exploits a deep network initialized with weights trained to detect tree crowns from UAV images. Pre-training on tree crowns and employing transfer learning allows the network to learn how to extract more relevant features for attack stage classification, and the limited task-specific data is effectively used to only train the shallow classification subnetwork. To overcome the data imbalance problem in the considered dataset, different data augmentation strategies were investigated, and affine warping was found to be the most effective for this purpose. Despite the challenges of inter-class overlap and intra-class non-homogeneity in the dataset, the proposed method achieved an average accuracy of 98.95%, significantly outperforming the baseline method.

Chapter 5

Conclusions

This thesis concludes by discussing the primary results for the three applications tackled in the preceding chapters. I first recap the key contributions to my primary goal in this thesis, applying computer vision techniques to solve practical forest monitoring problems with remote sensing (RS) data. In addition, I discuss the implications of each work individually and then briefly discuss the possibilities for future work.

5.1 Summary of Contributions

In this thesis, I first focused on generating high-quality orthomosaics from thermal drone imagery (Chapter 2). Traditional, single-modality orthomosaicking workflows often yield poor results when applied to drone-collected thermal images of forests due to the inherent low-resolution and low-contrast of such images. The necessary SfM process during orthomosaicking fails to accurately infer the 3D structure of the AOI, meaning that the later stages of orthomosaic generation also perform poorly. As a result, the generated thermal orthomosaics are of poor quality, failing to cover the entire area due to a large proportion of unaligned images. The areas that are present often have swirling distortions that muddle neighbouring trees. These issues render the generated orthomosaics unsuitable for downstream applications like ITCD. To solve this significant problem, I proposed a novel orthomosaicking workflow that bypasses the need for SfM on thermal images by instead using the SfM results of simultaneously

acquired RGB images for further processing stages. Once a 2.5D surface mesh of the AOI is constructed from the RGB images by ODM, thermal images need only be used to texture this mesh, and the final thermal orthomosaic is produced through a simple ortho-projection of the textured 2.5D model. Despite being taken simultaneously, objects do not initially appear in the same pixel locations for each RGB-thermal image pair. Therefore, to ensure the RGB-computed surface mesh is correctly textured using thermal images, the image pairs must first be properly co-registered. For this purpose, I proposed an ML-based strategy to accurately compute a single transformation matrix that precisely registers every image pair taken during a drone flight. Through extensive experimentation, a batched multi-resolution framework using the NGF loss function yielded the most precise registration results in terms of the MI metric for all five flight dates considered. Further, it was found that using a linear affine transformation matrix (i.e., rotation, scale, and skew) led to the best results regarding the visual appearance of the orthomosaic. The proposed workflow generates orthomosaics that are geometrically aligned, with objects appearing in the same pixel locations in both, as exemplified by a downstream ITCD task using an RGB-only detection model. It was further demonstrated that the proposed workflow correctly preserves radiometric information – temperature values were nearly identical for randomly selected locations in the individual thermal drone images and their corresponding locations in the generated thermal orthomosaic. The proposed integrated orthomosaicking workflow is applicable when both modalities are simultaneously acquired, which is possible with many commercially available multi-sensor instruments. The quality of the generated thermal orthomosaic is bounded by the quality of its generated RGB counterpart. Thus, standardized best practices should be followed during drone data acquisition to ensure the generation of a high-quality RGB orthomosaic and the necessary intermediate outputs. There must be sufficient overlap in both directions between successive images (at least 75%, but the higher, the better), and the drone flight speed should be slow enough to prevent motion blur. Weather conditions should be favourable as well. Too

much wind can destabilize the drone, leading to blurry images or deviations from the intended flight path. At the same time, raindrops on the camera lens may cause significant occlusion or artifacts in the images.

In Chapter 3, I shifted my focus to detecting individual tree crowns from overhead drone imagery. This chapter expands upon the straightforward ITCD example performed in the previous chapter by proposing a novel DL-based model (ShadowSense) that successfully uses RGB and thermal images to detect trees in challenging illumination conditions. In addition, I proposed a partially annotated, large-scale dataset (RT-Trees) based on the Cynthia cutblock acquisition for the experimental evaluation in this chapter and for facilitating future research for self-supervised ITCD in difficult lighting scenarios. The need for the proposed method and dataset was clear. The RGB-only detection model used in Chapter 2 fails to identify shorter trees hidden by the shadows of neighbouring taller trees (referred to as shadowed trees). Even manual inspection proves ineffective for delineating these tree crowns using only RGB imagery due to their inadequate visibility in these images. However, these crowns are apparent in the illumination-invariant thermal imagery. The need for the proposed method and dataset was clear. Therefore, ShadowSense exploits complementary information from both modalities to improve ITCD performance. During training, the RetinaNet backbone network pre-trained on RGB images is adapted to the thermal data distribution without any annotations through DAT using the foreground regions of the registered RGB-thermal image pairs in RT-Trees. Binary masks generated through a classic segmentation technique differentiate low-illumination areas (background) from brighter areas (foreground). Once the thermal backbone has been adapted to extract similar features as the RGB backbone for tree crowns in the foreground regions, the extracted features from both modalities in the background regions are fused in a weighted manner and passed to the detection heads to produce bounding box outputs. This combined model has the advantages of (1) successfully identifying completely shadowed trees and (2) improving the precision of bounding box limits for trees partially hidden in

shadows. Quantitative results on the manually annotated test set of RT-Trees showed that the proposed model detected tree crowns with a higher precision and recall score than the baseline RGB-only detector and existing multimodal methods based on UDA or early image-level fusion. Importantly, the number of shadowed trees detected was significantly higher using the proposed RGB-thermal detection model.

The tree crowns detected in Chapter 3 are useful for determining specific properties of individual trees. To demonstrate this, in Chapter 4 I tackled the problem of classifying images of individual tree crowns into one of four bark beetle attack stages. Once enough bark beetles infest a pine tree and overwhelm its natural defences, the host tree's crown gradually fades from a healthy green to yellow, red, and then a lifeless grey (i.e., the pine needles fall off). Manual surveys have long been the standard approach for mapping bark beetle infestations using these visual symptoms. However, this can quickly become arduous as the area considered grows, prompting the need for an automated classification model. Although this may seem like a simple colour classification task, overlapping crowns and the gradual, non-homogenous nature of crown discoloration due to variations in host tree defensive responses pose a considerable challenge. DL models can be trained to learn general feature representations to account for such biological and physical characteristics for successful classification. However, there is a scarcity of labelled data for this task to train a DL model from scratch using supervised learning. Therefore, I employed a transfer learning strategy to fine-tune a pre-trained ITCD model for this task (i.e., the one used in Section 2.3.5 and the baseline in Chapter 3). The ResNet-50 and FPN backbone weights were initialized from the pre-trained model, while the detection heads were replaced with a shallow classification subnetwork specific to this task. The model was then trained end-to-end using the limited labelled data. Additionally, augmentation was employed to balance the number of samples from each class and simultaneously increase the total number of training data samples. Experimental results on an existing bark beetle attack dataset demonstrated the superiority of the proposed DL

model over the existing cellular automaton baseline method and various classical ML methods. Different augmentation strategies were analyzed, and affine warping yielded near-perfect classification accuracy, improving on the baseline results for this dataset by nearly 10%. Once the distribution of trees infested and killed in the previous years is mapped using the proposed classification model, infestation trends can be predicted using the proximity of red-attack trees. The presence of bark beetles can be confirmed through manual inspection before felling or burning (a common control method). Manual inspection involves searching for external-bole symptoms like pitch tubes or boring dust and generally provides more conclusive evidence of bark beetle infestations than crown fading, which other factors like drought or different pests may cause. Using overhead drone imagery alone is not as effective in distinguishing between these stressors. However, the search area for recently infested trees can be significantly reduced thanks to the proposed model since emerging bark beetles are known to engage in short-range dispersal preferentially.

5.2 Implications

With the pace of innovation in the computer vision field, RS applications for FHM can be solved like never before. As I showed in this thesis, ML and DL, in particular, can be applied to various parts of the RS pipeline, from preprocessing (orthomosaicking) to downstream analysis (attack stage classification).

A particularly significant component of this thesis was the utilization of multi-modal drone imagery. Despite the challenges associated with multiple data sources, such as differences in contrast, resolution, and misalignment, complementary information can be combined to solve problems more effectively than using any one modality alone. I demonstrated this for two of the applications considered. In the case of orthomosaicking, intermediate results from RGB images were used to enhance the thermal processing and provide a more comprehensive representation of the AoI. For ITCD, I showed that fusing features from both modalities improves on RGB-only

detection by effectively identifying trees hidden in shadows.

An overall RS pipeline comprising the works proposed in the thesis can be directly employed to solve practical problems like monitoring bark beetle attacks. For instance, forest managers can collect RGB-thermal drone imagery, generate two high-quality orthomosaics of the AOI, extract tree crowns, and classify the bark beetle attack stage of each to help determine the distribution of infested trees. Alternatively, the three components can be independently applied to the corresponding parts of existing RS pipelines. High-quality thermal orthomosaics alone can be used to determine fire fronts and rate of spread for forest fire monitoring without the need to detect or classify tree crowns. Similarly, the proposed shadow-agnostic ITCD model can be applied to drone images directly rather than first generating an orthomosaic, or the attack stage classifier can be used on manually extracted tree crowns, depending on the requirements of forest managers.

5.3 Future Work

The methods proposed in this thesis lay the foundation for future research in several possible directions. For instance, the open-source nature of the orthomosaicking tool described in Chapter 2 inherently allows for continuous improvement. One such improvement could involve integrating specific postprocessing techniques on the generated orthomosaics, such as thermal drift correction for handling variations in thermal sensor readings as the instrument warms up during the flight and thus produce more robust orthomosaics. Another improvement could be allowing diffeomorphic transformations to achieve more precise registration of images that may not be completely undistorted. An important consideration would be to do this in a way that doesn't interfere with the stitching algorithm so that individual trees don't get duplicated or omitted. Over time, I believe the tool will continue to grow as a valuable resource for the RS community. Furthermore, as long as simultaneously acquired RGB images are available, the underlying proposed method can be used for applications

other than forest monitoring, e.g., urban or agricultural monitoring, or co-registered modalities other than thermal, e.g., multispectral.

The work presented in the other two chapters can also be extended in future research projects. The shadow-agnostic ITCD model proposed in Chapter 3 successfully detects tree crowns from forest images, and the deep classification model presented in Chapter 4 them into four visible bark beetle attack stages. Although this can be a valuable tool for understanding the temporal and spatial distribution of beetle infestations during previous years, by the time the red attack stage sets in, it is usually *after* the next generation of beetles hatches and disperses, making it challenging to control mass outbreaks effectively. Therefore, a more useful tool could be one that can identify the presence of bark beetles *before* the onset of visual crown fading, i.e., green attack identification. This remains a difficult task due to factors such as variations in the biological response of host trees and weather conditions during data collection [140]. However, a classification model that can accomplish this task would be exceedingly valuable for the FHM community. One possible solution for this task could be to use thermal images to identify signal changes occurring before the onset of visual symptoms as an indicator of bark beetle presence. The orthomosaicking and registration methods presented in Chapter 2 can be effectively applied to this solution. The detection model proposed in Chapter 3 that successfully detects shorter shadowed trees will be especially helpful for green attack classification, considering that bark beetles preferentially attack these younger, weaker trees during their endemic population stage.

Bibliography

[1] H. Aasen, E. Honkavaara, A. Lucieer, and P. Zarco-Tejada, "Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correction workflows," *Remote Sensing*, vol. 10, no. 7, p. 1091, Jul. 2018. DOI: 10.3390/rs10071091.

[2] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[3] I. B. Akkaya, F. Altinel, and U. Halici, "Self-training guided adversarial domain adaptation for thermal imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4322–4331.

[4] G. Bareth and G. Waldhoff, "GIS for mapping vegetation," in *Comprehensive Geographic Information Systems*, Elsevier, 2018, pp. 1–27. DOI: 10.1016/b978-0-12-409548-9.09636-6.

[5] M. Beloiu, L. Heinzmann, N. Rehush, A. Gessler, and V. C. Griess, "Individual tree-crown detection and species identification in heterogeneous forests using aerial RGB imagery and deep learning," *Remote Sensing*, vol. 15, no. 5, p. 1463, Mar. 2023. DOI: 10.3390/rs15051463.

[6] K. O. Bergmüller and M. C. Vanderwel, "Predicting tree mortality using spectral indices derived from multispectral UAV imagery," *Remote Sensing*, vol. 14, no. 9, p. 2195, May 2022. DOI: 10.3390/rs14092195.

[7] S. Beucher and F. Meyer, "The morphological approach to segmentation: The watershed transformation," in *Mathematical Morphology in Image Processing*, E. Dougherty, Ed., Marcel Dekker Inc., New York, 1993, pp. 433–481.

[8] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[9] J. Blanco-Sacristán *et al.*, "UAV RGB, thermal infrared and multispectral imagery used to investigate the control of terrain on the spatial distribution of dryland biocrust," *Earth Surface Processes and Landforms*, vol. 46, no. 12, pp. 2466–2484, Aug. 2021. DOI: 10.1002/esp.5189.

[10] K. P. Bleiker and B. H. V. Hezewijk, "Flight period of mountain pine beetle (coleoptera: Curculionidae) in its recently expanded range," *Environmental Entomology*, vol. 45, no. 6, pp. 1561–1567, Sep. 2016. DOI: 10.1093/ee/nvw121. [Online]. Available: https://doi.org/10.1093/ee/nvw121.

[11] K. P. Bleiker *et al.*, "Risk assessment of the threat of mountain pine beetle to canada's boreal and eastern pine forests.," *Canadian Council of Forest Ministers, Ottawa, Ontario.*, p. 65, 2019.

[12] D. C. Brown, "Decentering distortion of lenses," *Photogrammetric Engineering and Remote Sensing*, vol. 32, no. 3, pp. 444–462, 1966.

[13] A. L. Carroll, T. L. Shore, and L. Safranyik, *Direct control: theory and practice*. Victoria, British Columbia: Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, British Columbia, 2006, ch. 6, pp. 155–172.

[14] D. Cernea, *Openmvs: Open multiple view stereovision.* Available online: https://github.com/cdcseacave/openMVS/.

[15] A. J. Chadwick *et al.*, "Automatic delineation and height measurement of regenerating conifer crowns under leaf-off conditions using UAV imagery," *Remote Sensing*, vol. 12, no. 24, p. 4104, Dec. 2020. DOI: 10.3390/rs12244104.

[16] J. Chen, J. Wei, and R. Li, "Targan: Target-aware generative adversarial networks for multi-modality medical image translation," 2021. arXiv: 2105.08993 [eess.IV].

[17] S. Dandrifosse, A. Carlier, B. Dumont, and B. Mercatoris, "Registration and fusion of close-range multimodal wheat images in field conditions," *Remote Sensing*, vol. 13, no. 7, p. 1380, Apr. 2021. DOI: 10.3390/rs13071380.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[19] M. Dillen, M. Vanhellemont, P. Verdonckt, W. H. Maes, K. Steppe, and K. Verheyen, "Productivity, stand dynamics and the selection effect in a mixed willow clone short rotation coppice plantation," *Biomass and Bioenergy*, vol. 87, pp. 46–54, Apr. 2016. DOI: 10.1016/j.biombioe.2016.02.013.

[20] A. Duarte, N. Borralho, P. Cabral, and M. Caetano, "Recent advances in forest insect pests and diseases monitoring using UAV-based data: A systematic review," *Forests*, vol. 13, no. 6, p. 911, Jun. 2022. DOI: 10.3390/f13060911.

[21] S. Ecke *et al.*, "UAV-based forest health monitoring: A systematic review," *Remote Sensing*, vol. 14, no. 13, p. 3205, 2022. DOI: 10.3390/rs14133205.

[22] N. Erbilgin *et al.*, "Combined drought and bark beetle attacks deplete non-structural carbohydrates and promote death of mature pine trees," *Plant, Cell & Environment*, vol. 44, no. 12, pp. 3866–3881, 2021. DOI: https://doi.org/10.1111/pce.14197.

[23] G. Evangelidis and E. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, Oct. 2008. DOI: 10.1109/tpami.2008.113.

[24] S. A. Fakhri and H. Latifi, "A consumer grade UAV-based framework to estimate structural attributes of coppice and high oak forest stands in semi-arid regions," *Remote Sensing*, vol. 13, no. 21, p. 4367, Oct. 2021. DOI: 10.3390/rs13214367.

[25] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision*, vol. 85, no. 1, pp. 35–57, 2009. DOI: 10.1007/s11263-009-0243-z.

[26] Fujimoto *et al.*, "An end to end process development for UAV-SfM based forest monitoring: Individual tree detection, species classification and carbon dynamics simulation," *Forests*, vol. 10, no. 8, p. 680, 2019. DOI: 10.3390/f10080680.

[27] L. Gan, C. Lee, and S.-J. Chung, "Unsupervised RGB-to-thermal domain adaptation via multi-domain attention network," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, May 2023. DOI: 10.1109/icra48891.2023.10160872.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 1180–1189. [Online]. Available: https://proceedings.mlr.press/v37/ganin15.html.

[29] P. Gao, T. Tian, T. Zhao, L. Li, N. Zhang, and J. Tian, "GF-detection: Fusion with GAN of infrared and visible images for vehicle detection at nighttime," *Remote Sensing*, vol. 14, no. 12, p. 2771, 2022. DOI: 10.3390/rs14122771.

[30] N. Guimarães, L. Pádua, P. Marques, N. Silva, E. Peres, and J. J. Sousa, "Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities," *Remote Sensing*, vol. 12, no. 6, p. 1046, Mar. 2020. DOI: 10.3390/rs12061046.

[31] L. Guo, S. Huang, D. Liu, H. Cheng, and B. Wen, "Shadowformer: Global context helps image shadow removal," 2023. arXiv: 2302.01650 [cs.CV].

[32] Q. Guo, W. Zhou, J. Lei, and L. Yu, "TSFNet: Two-stage fusion network for RGB-t salient object detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1655–1659, 2021. DOI: 10.1109/lsp.2021.3102524.

[33] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115, 2017.

[34] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, Springer Berlin Heidelberg, 2006, pp. 726–733. DOI: 10.1007/11866763_89.

[35] B. C. Hall, "An elementary introduction to groups and representations," 2000. arXiv: math-ph/0005032 [math-ph].

[36] B. C. Hall, *Lie Groups, Lie Algebras, and Representations*. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-13467-3.

[37] R. Hall, G. Castilla, J. White, B. Cooke, and R. Skakun, "Remote sensing of forest pest damage: A review and lessons learned from a canadian perspective," *The Canadian Entomologist*, vol. 148, no. S1, S296–S356, May 2016. DOI: 10.4039/tce.2016.11.

[38] Q. Han and C. Jung, "Deep selective fusion of visible and near-infrared images using unsupervised u-net," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022. DOI: 10.1109/tnnls.2022.3142780.

[39] S. N. H. S. Hanapi, S. A. A. Shukor, and J. Johari, "A review on remote sensing-based method for tree detection and delineation," *IOP Conference Series: Materials Science and Engineering*, vol. 705, no. 1, p. 012 024, Nov. 2019. DOI: 10.1088/1757-899x/705/1/012024.

[40] Hanusch, Thomas, "Texture mapping and true orthophoto generation of 3d objects," en, Ph.D. dissertation, ETH Zurich, 2010. DOI: 10.3929/ETHZ-A-006194643.

[41] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Second. New York, NY, USA: Cambridge University Press, ISBN: 0521540518, 2004, ch. 2, pp. 25–64.

[42] W. Hartmann, S. Tilch, H. Eisenbeiss, and K. Schindler, "DETERMINATION OF THE UAV POSITION BY AUTOMATIC PROCESSING OF THERMAL IMAGES," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXIX-B6, pp. 111–116, Jul. 2012. DOI: 10.5194/isprsarchives-xxxix-b6-111-2012.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

[44] H. Hoffmann, H. Nieto, R. Jensen, R. Guzinski, P. Zarco-Tejada, and T. Friborg, "Estimating evaporation with thermal UAV data and two-source energy balance models," *Hydrology and Earth System Sciences*, vol. 20, no. 2, pp. 697–713, Feb. 2016. DOI: 10.5194/hess-20-697-2016.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.

[46] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 2261–2269.

[47] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[48] K. Iizuka *et al.*, "Visualizing the spatiotemporal trends of thermal characteristics in a peatland plantation forest in indonesia: Pilot test using unmanned aerial systems (UASs)," *Remote Sensing*, vol. 10, no. 9, p. 1345, Aug. 2018. DOI: 10.3390/rs10091345.

[49] F. Javadnejad, D. T. Gillins, C. E. Parrish, and R. K. Slocum, "A photogrammetric approach to fusing natural colour and thermal infrared UAS imagery in 3d point cloud generation," *International Journal of Remote Sensing*, vol. 41, no. 1, pp. 211–237, Jul. 2019. DOI: 10.1080/01431161.2019.1641241.

[50] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 3489–3497.

[51] F. Jiang, A. R. Smith, M. Kutia, G. Wang, H. Liu, and H. Sun, "A modified KNN method for mapping the leaf area index in arid and semi-arid areas of china," *Remote Sensing*, vol. 12, no. 11, p. 1884, Jun. 2020. DOI: 10.3390/rs12111884.

[52] S. Junttila *et al.*, "Multispectral imagery provides benefits for mapping spruce tree decline due to bark beetle infestation when acquired late in the season," *Remote Sensing*, vol. 14, no. 4, p. 909, Feb. 2022. DOI: 10.3390/rs14040909.

[53] R. Kapil, G. Castilla, S. M. Marvasti-Zadeh, D. Goodsman, N. Erbilgin, and N. Ray, "Orthomosaicking thermal drone images of forests via simultaneously acquired RGB images," *Remote Sensing*, vol. 15, no. 10, p. 2653, 2023. DOI: 10.3390/rs15102653.

[54] R. Kapil, S. M. Marvasti-Zadeh, D. Goodsman, N. Ray, and N. Erbilgin, "Classification of bark beetle-induced forest tree mortality using deep learning," 2022. arXiv: 2207.07241 [cs.CV].

[55] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, Jul. 2013, ISSN: 0730-0301. DOI: 10.1145/2487228.2487237. [Online]. Available: https://doi.org/10.1145/2487228.2487237.

[56] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[57] M. Kieu, A. D. Bagdanov, M. Bertini, and A. del Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 546–562. DOI: 10.1007/978-3-030-58542-6_33.

[58] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon, "MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6497–6504, 2021. DOI: 10.1109/lra.2021.3093652.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. arXiv: 1412.6980 [cs.LG].

[60] L. Konig and J. Ruhaak, "A fast and accurate parallel algorithm for non-linear image registration using normalized gradient fields," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Apr. 2014. DOI: 10.1109/isbi.2014.6867937.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 2, 2012, pp. 1097–1105.

[62] A. Li, D. Ye, E. Lyu, S. Song, M. Q.-H. Meng, and C. W. de Silva, "RGB-thermal fusion network for leakage detection of crude oil transmission pipes," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2019. DOI: 10.1109/robio49542.2019.8961733.

[63] H. Li, W. Ding, X. Cao, and C. Liu, "Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing," *Remote Sensing*, vol. 9, no. 5, p. 441, May 2017. DOI: 10.3390/rs9050441.

[64] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for RGB-thermal perception tasks," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 4060–4067, Jul. 2023. DOI: 10.1109/lra.2023.3272269.

[65] G. Liao, W. Gao, G. Li, J. Wang, and S. Kwong, "Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7646–7661, Nov. 2022. DOI: 10.1109/tcsvt.2022.3184840.

[66] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017. DOI: 10.1109/cvpr.2017.106.

[67] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[68] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 105–119, 2022. DOI: 10.1109/tcsvt.2021.3056725.

[69] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," 2022. arXiv: 2211.10960 [cs.CV].

[70] J. Liu *et al.*, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022. DOI: 10.1109/cvpr52688.2022.00571.

[71] Y. Liu, C. Zheng, X. Liu, Y. Tian, J. Zhang, and W. Cui, "Forest fire monitoring method based on UAV visual and infrared image fusion," *Remote Sensing*, vol. 15, no. 12, p. 3173, Jun. 2023. DOI: 10.3390/rs15123173.

[72] A. López, J. M. Jurado, C. J. Ogayar, and F. R. Feito, "A framework for registering UAV-based imagery for crop-tracking in precision agriculture," *International Journal of Applied Earth Observation and Geoinformation*, vol. 97, p. 102 274, May 2021. DOI: 10.1016/j.jag.2020.102274.

[73] A. López, C. J. Ogayar, J. M. Jurado, and F. R. Feito, "Efficient generation of occlusion-aware multispectral and thermographic point clouds," *Computers and Electronics in Agriculture*, vol. 207, p. 107 712, Apr. 2023. DOI: 10.1016/j.compag.2023.107712.

[74] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. arXiv: 1711.05101 [cs.LG].

[75] D. G. Lowe, "Object recognition from local scale-invariant features," in *seventh IEEE international conference on computer vision*, IEEE, vol. 2, 1999, pp. 1150–1157.

[76] F. Luo, Y. Li, G. Zeng, P. Peng, G. Wang, and Y. Li, "Thermal infrared image colorization for nighttime driving scenes with top-down guided attention," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 808–15 823, 2022. DOI: 10.1109/tits.2022.3145476.

[77] C. Lyu, P. Heyer, B. Goossens, and W. Philips, "An unsupervised transfer learning framework for visible-thermal pedestrian detection," *Sensors*, vol. 22, no. 12, p. 4416, 2022. DOI: 10.3390/s22124416.

[78] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, Apr. 1997. DOI: 10.1109/42.563664.

[79] W. Maes, A. Huete, and K. Steppe, "Optimizing the processing of UAV-based thermal imagery," *Remote Sensing*, vol. 9, no. 5, p. 476, May 2017. DOI: 10.3390/rs9050476.

[80] W. Maes *et al.*, "Can UAV-based infrared thermography be used to study plant-parasite interactions between mistletoe and eucalypt trees?" *Remote Sensing*, vol. 10, no. 12, p. 2062, Dec. 2018. DOI: 10.3390/rs10122062.

[81] S. Manfreda *et al.*, "On the use of unmanned aerial systems for environmental monitoring," *Remote Sensing*, vol. 10, no. 4, p. 641, Apr. 2018. DOI: 10.3390/rs10040641.

[82] Mapillary, *Mapillary-opensfm. an open-source structure from motion library that lets you build 3d models from images*, Available online: https://opensfm.org/.

[83] S. M. Marvasti-Zadeh, D. Goodsman, N. Ray, and N. Erbilgin, "Early detection of bark beetle attack using remote sensing and machine learning: A review," Oct. 2022. DOI: 10.48550/arxiv.2210.03829. [Online]. Available: https://arxiv.org/abs/2210.03829.

[84] S. M. Marvasti-Zadeh, D. Goodsman, N. Ray, and N. Erbilgin, "Crown-CAM: Interpretable visual explanations for tree crown detection in aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[85] E. Maset, A. Fusiello, F. Crosilla, R. Toldo, and D. Zorzetto, "PHOTOGRAMMETRIC 3D BUILDING RECONSTRUCTION FROM THERMAL IMAGES," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W3, pp. 25–32, Aug. 2017. DOI: 10.5194/isprs-annals-iv-2-w3-25-2017.

[86] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 423–433.

[87] L. Merino, F. Caballero, J. R. Martínez-de-Dios, I. Maza, and A. Ollero, "An unmanned aircraft system for automatic forest fire monitoring and measurement," *Journal of Intelligent Robotic Systems*, vol. 65, no. 1-4, pp. 533–548, Aug. 2011. DOI: 10.1007/s10846-011-9560-x.

[88] R. Minařík, J. Langhammer, and T. Lendzioch, "Detection of bark beetle disturbance at tree level using UAS multispectral imagery and deep learning," *Remote Sensing*, vol. 13, no. 23, 2021.

[89] F. Moradi, F. D. Javan, and F. Samadzadegan, "Potential evaluation of visible-thermal UAV image fusion for individual tree detection based on convolutional neural network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 103 011, 2022. DOI: 10.1016/j.jag.2022.103011.

[90] F. Munir, S. Azam, and M. Jeon, "SSTN: Self-supervised domain adaptation thermal object detection for autonomous driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021. DOI: 10.1109/iros51168.2021.9636353.

[91] A. Namburu, P. Selvaraj, S. Mohan, S. Ragavanantham, and E. T. Eldin, "Forest fire identification in UAV imagery using x-MobileNet," *Electronics*, vol. 12, no. 3, p. 733, Feb. 2023. DOI: 10.3390/electronics12030733.

[92] A. Nan, M. Tennant, U. Rubin, and N. Ray, "Drmime: Differentiable mutual information and matrix exponential for multi-resolution image registration," in *Third Conference on Medical Imaging with Deep Learning*, T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds., ser. Proceedings of Machine Learning Research, vol. 121, PMLR, Jul. 2020, pp. 527–543. [Online]. Available: https://proceedings.mlr.press/v121/nan20a.html.

[93] H. T. Nguyen, M. L. Lopez Caceres, K. Moritake, S. Kentsch, H. Shu, and Y. Diez, "Individual sick fir tree (abies mariesii) identification in insect infested forests by means of UAV images and deep learning," *Remote Sensing*, vol. 13, no. 2, 2021.

[94] O. F. Niemann and F. Visintini, "Assessment of potential for remote sensing detection of bark beetle-infested areas during green attack : A literature review," *Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre*, 2005.

[95] A. Olejnik *et al.*, "The use of unmanned aerial vehicles in remote sensing systems," *Sensors*, vol. 20, no. 7, p. 2003, Apr. 2020. DOI: 10.3390/s20072003.

[96] M. Onishi and T. Ise, "Explainable identification and mapping of trees using UAV RGB image and deep learning," *Scientific Reports*, vol. 11, no. 1, 2021. DOI: 10.1038/s41598-020-79653-9.

[97] OpenDroneMap Authors, *Odm – a command line toolkit to generate maps, point clouds, 3d models and dems from drone, balloon or kite images.* OpenDroneMap/ODM GitHub Page 2020. Available online: https://github.com/OpenDroneMap/ODM.

[98] T. A. Ouattara, V.-C. J. Sokeng, I. C. Zo-Bi, K. F. Kouamé, C. Grinand, and R. Vaudry, "Detection of forest tree losses in côte d'ivoire using drone aerial images," *Drones*, vol. 6, no. 4, p. 83, Mar. 2022. DOI: 10.3390/drones6040083.

[99] P. Oza, V. A. Sindagi, V. V. Sharmini, and V. M. Patel, "Unsupervised domain adaptation of object detectors: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–24, 2023. DOI: 10.1109/tpami.2022.3217046.

[100] G. Pasqualino, A. Furnari, G. Signorello, and G. M. Farinella, "An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites," *Image and Vision Computing*, vol. 107, p. 104 098, 2021. DOI: 10.1016/j.imavis.2021.104098.

[101] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[102] M. Pause *et al.*, "In situ/remote sensing integration to assess forest health—a review," *Remote Sensing*, vol. 8, no. 6, p. 471, Jun. 2016. DOI: 10.3390/rs8060471.

[103] G. Penney, J. Weese, J. Little, P. Desmedt, D. Hill, and D. hawkes, "A comparison of similarity measures for use in 2-d-3-d medical image registration," *IEEE Transactions on Medical Imaging*, vol. 17, no. 4, pp. 586–595, 1998. DOI: 10.1109/42.730403.

[104] K. M. Potter and B. L. Conkling, "Forest health monitoring: National status, trends, and analysis 2021," Tech. Rep., May 2022. DOI: 10.2737/srs-gtr-266.

[105] K. Ribeiro-Gomes, D. Hernández-López, J. Ortega, R. Ballesteros, T. Poblete, and M. Moreno, "Uncooled thermal camera calibration and optimization of the photogrammetry process for UAV applications in agriculture," *Sensors*, vol. 17, no. 10, p. 2173, Sep. 2017. DOI: 10.3390/s17102173.

[106] S. Ruder, "An overview of gradient descent optimization algorithms," 2017. arXiv: 1609.04747 [cs.LG].

[107] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986. DOI: 10.1038/323533a0.

[108] A. Safonova, S. Tabik, D. Alcaraz-Segura, A. Rubtsov, Y. Maglinets, and F. Herrera, "Detection of fir trees (abies sibirica) damaged by the bark beetle in unmanned aerial vehicle images with deep learning," *Remote Sensing*, vol. 11, no. 6, 2019.

[109] L. Safranyik and A. L. Carroll, "The biology and epidemiology of the mountain pine beetle in lodgepole pine forests," *The Mountain Pine Beetle: A Synthesis of Its Biology, Management and Impacts on Lodgepole Pine*, pp. 3–66, 2006.

[110] L. Safranyik, D. A. Linton, R. Silversides, and L. H. McMullen, "Dispersal of released mountain pine beetles under the canopy of a mature lodgepole pine stand," *Journal of Applied Entomology*, vol. 113, no. 1-5, pp. 441–450, 1992.

[111] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[112] S. Saleem and A. Bais, "Visible spectrum and infra-red image matching: A new method," *Applied Sciences*, vol. 10, no. 3, p. 1162, Feb. 2020. DOI: 10.3390/app10031162.

[113] S. Schaeffer *et al.*, "Detection of bark beetle infestation in drone imagery via thresholding cellular automata," *Journal of Applied Remote Sensing*, vol. 15, no. 1, pp. 1–20, 2021.

[114] J. Sedano-Cibrián, R. Pérez-Álvarez, J. M. de Luis-Ruiz, R. Pereda-García, and B. R. Salas-Menocal, "Thermal water prospection with UAV, low-cost sensors and GIS. application to the case of la hermida," *Sensors*, vol. 22, no. 18, p. 6756, Sep. 2022. DOI: 10.3390/s22186756.

[115] S. Shen, "Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1901–1914, May 2013. DOI: 10.1109/tip.2013.2237921.

[116] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–14.

[117] A. Sledz, J. Unger, and C. Heipke, "THERMAL IR IMAGING: IMAGE QUALITY AND ORTHOPHOTO GENERATION," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-1, pp. 413–420, Sep. 2018. DOI: 10.5194/isprs-archives-xlii-1-413-2018.

[118] M. Smigaj, R. Gaulton, S. L. Barr, and J. C. Suárez, "UAV-BORNE THERMAL IMAGING FOR FOREST HEALTH MONITORING: DETECTION OF DISEASE INDUCED CANOPY TEMPERATURE INCREASE," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-3/W3, pp. 349–354, Aug. 2015. DOI: 10.5194/isprsarchives-xl-3-w3-349-2015.

[119] P. J. Soille and M. M. Ansoult, "Automated basin delineation from digital elevation models using mathematical morphology," *Signal Processing*, vol. 20, no. 2, pp. 171–182, 1990, ISSN: 0165-1684. DOI: https://doi.org/10.1016/0165-1684(90)90127-K. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016516849090127K.

[120] S. Song *et al.*, "Deep domain adaptation based multi-spectral salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 128–140, 2022. DOI: 10.1109/tmm.2020.3046868.

[121] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019. DOI: 10.1109/lra.2019.2904733.

[122] T. D. Swaef *et al.*, "Applying RGB- and thermal-based vegetation indices from UAVs for high-throughput field phenotyping of drought tolerance in forage grasses," *Remote Sensing*, vol. 13, no. 1, p. 147, Jan. 2021. DOI: 10.3390/rs13010147.

[123] R. Szeliski, *Computer Vision*. Springer London, 2011, ch. 3, pp. 107–190. DOI: 10.1007/978-1-84882-935-0.

[124] R. Szeliski, *Computer Vision Algorithms and Applications 2nd Edition*. Springer London, 2021, ch. 7, pp. 483–491, ISBN: 3030343715.

[125] A. Toet, *TNO image fusion dataset*, 2014. DOI: 10.6084/M9.FIGSHARE.1008029.V2.

[126] T. P. Truong, M. Yamaguchi, S. Mori, V. Nozick, and H. Saito, "Registration of RGB and thermal point clouds generated by structure from motion," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Oct. 2017. DOI: 10.1109/iccvw.2017.57.

[127] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for RGB-thermal salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 5678–5691, 2021. DOI: 10.1109/tip.2021.3087412.

[128] S. Ullman, "The interpretation of structure from motion," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979.

[129] V. Vidit and M. Salzmann, "Attention-based domain adaptation for single-stage detectors," *Machine Vision and Applications*, vol. 33, no. 5, Jul. 2022. DOI: 10.1007/s00138-022-01320-y.

[130] V. VS, D. Poster, S. You, S. Hu, and V. M. Patel, "Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 1412–1423.

[131] S. van der Walt *et al.*, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, e453, Jun. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.453. [Online]. Available: https://doi.org/10.7717/peerj.453.

[132] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," 2022. arXiv: 2205.11876 [cs.CV].

[133] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-guided fusion network for RGB-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2949–2961, 2022. DOI: 10.1109/tcsvt.2021.3099120.

[134] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1788–1797. DOI: 10.1109/CVPR.2018.00192.

[135] Z. Wang, W. Shao, Y. Chen, J. Xu, and X. Zhang, "Infrared and visible image fusion via interactive compensatory attention adversarial learning," *IEEE Transactions on Multimedia*, pp. 1–13, 2023. DOI: 10.1109/tmm.2022.3228685.

[136] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks," *Remote Sensing*, vol. 11, no. 11, p. 1309, Jun. 2019. DOI: 10.3390/rs11111309.

[137] B. G. Weinstein, S. Marconi, and E. White, "Training data for the NEON tree evaluation benchmark," *Zenodo 10.5281/zenodo.5912107*, Jan. 2022.

[138] K. Whitehead and C. H. Hugenholtz, "Remote sensing of the environment with small unmanned aircraft systems (UASs), part 1: A review of progress and challenges," *Journal of Unmanned Vehicle Systems*, vol. 02, no. 03, pp. 69–85, Sep. 2014. DOI: 10.1139/juvs-2014-0006.

[139] P. Wilmott, *The mathematics of financial derivatives, a student introduction*. Cambridge University Press, 1995, p. 317, ISBN: 0521496993.

[140] M. A. Wulder, J. C. White, A. L. Carroll, and N. C. Coops, "Challenges for the operational detection of mountain pine beetle green attack with remote sensing," *The Forestry Chronicle*, vol. 85, no. 1, pp. 32–38, 2009. DOI: 10.5558/tfc85032-1. eprint: https://doi.org/10.5558/tfc85032-1. [Online]. Available: https://doi.org/10.5558/tfc85032-1.

[141] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022. DOI: 10.1109/tpami.2020.3012548.

[142] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 484–12 491, 2020. DOI: 10.1609/aaai.v34i07.6936.

[143] S. Yahyanejad and B. Rinner, "A fast and mobile system for registration of low-altitude visual and thermal aerial images using multiple small-scale UAVs," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 104, pp. 189–202, Jun. 2015. DOI: 10.1016/j.isprsjprs.2014.07.015.

[144] Y. Yang and X. Lee, "Four-band thermal mosaicking: A new method to process infrared thermal imagery of urban landscapes from UAV flights," *Remote Sensing*, vol. 11, no. 11, p. 1365, Jun. 2019. DOI: 10.3390/rs11111365.

[145] Y. Yang and N. Ray, "Foreground-focused domain adaption for object detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021. DOI: 10.1109/icpr48806.2021.9412906.

[146] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. WACV*, 2021, pp. 3258–3267.

[147] A. Zakrzewska and D. Kopeć, "Remote sensing of bark beetle damage in norway spruce individual tree canopies using thermal infrared and airborne laser scanning data fusion," *Forest Ecosystems*, vol. 9, p. 100 068, 2022. DOI: 10.1016/j.fecs.2022.100068.

[148] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1804–1818, 2021. DOI: 10.1109/tcsvt.2020.3014663.

[149] X. Zhang, P. Ye, and G. Xiao, "Vifb: A visible and infrared image fusion benchmark," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 468–478, 2020.

[150] Z. Zhang and L. Zhu, "A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications," *Drones*, vol. 7, no. 6, p. 398, Jun. 2023. DOI: 10.3390/drones7060398.

[151] H. Zhao, J. Morgenroth, G. Pearse, and J. Schindler, "A systematic review of individual tree crown detection and delineation with convolutional neural networks (CNN)," *Current Forestry Reports*, 2023. DOI: 10.1007/s40725-023-00184-3.

[152] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 955–13 965.

[153] Z. Zhao, Y. Guo, H. Shen, and J. Ye, "Adaptive object detection with dual multi-label prediction," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 54–69. DOI: 10.1007/978-3-030-58604-1_4.

[154] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[155] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022. DOI: 10.1109/tcsvt.2021.3077058.

Appendix A: RT-Trees Dataset Additional Information

In this appendix, the collection, pre-processing, and annotation procedures employed for the RT-Trees dataset is described. Furthermore, instances of challenging scenarios within the proposed dataset are showcased to provide a broader perspective.

For this dataset, the Cynthia cutblock data introduced in Section 2.2 was expanded through nine additional drone flights conducted between September and November 2022 (14 flights in total) with the same drone and sensor setup. Specifically, a multi-camera DJI H20T sensor instrument was used to simultaneously acquire wide-angle RGB images of 4056×3040 pixels with an $82.9°$ FOV and thermal images in the 8-14μm spectral band of 640×512 pixels with $40.6°$ FOV. The RGB camera uses a 1/2.3C̈MOS (12 MP) sensor, while the thermal camera uses an uncooled VOx microbolometer sensor. The sensor instrument was mounted to a Matrice 300 RTK drone and successive image pairs were captured with an 80% front and 75% side overlap in the thermal images via a fixed flight path identical to the path used for the original five flights.

The image pairs from all 14 flights were preprocessed as follows. Thermal images were first upscaled through bilinear interpolation to 1500×1000, while RGB images were centre-cropped to 1500×1000 pixels to discard edge distortions. This cropping size also ensures that both images in a pair display roughly the same amount of area. Each RGB-thermal pair was precisely co-registered using the normalized gradient fields-based workflow described in [53]. Fig. A.7 shows an example pair of 1500×1000

Table A.1: **RT-Trees Dataset Information by Flight Date**. All dates are from the year 2022. Information about the flight, lighting and weather conditions, and number of images is listed. Approximately 70% of the raw image pairs captured for a given date are sampled for the training set based on GPS location (see Fig. A.2), and then divided into six 500×500 patches. From the August 30 data, 63 images are taken for testing and 10 for validation, hence the total number of image pairs in RT-Trees is 49879.

Flight Date	**Time of First Capture (24h)**	**Flight Duration (min)**	**Sun Elevation (°)**	**Sun Azimuth (°)**	**Air Temperature (°C)**	**Raw Image Pairs Captured**	**500×500 Patches in Training Set**
July 20	11:04	27	44.77	120.15	20.3	827	3582
July 26	10:18	28	37.51	109.32	20.8	828	3588
August 9	10:16	28	34.45	111.71	19.8	820	3552
August 17	12:15	33	46.12	147.29	24.5	825	3570
August 30	11:21	31	37.26	134.21	25.4	814	3516
September 9	11:40	32	36.08	142.39	14.0	824	3570
September 15	11:00	27	30.20	133.15	17.5	825	3582
September 23	11:14	28	29.12	139.04	14.7	820	3552
October 4	11:13	31	25.43	141.56	16.8	820	3558
October 6	15:16	27	27.22	210.10	9.8	808	3498
October 7	19:02	27	0.26	261.26	2.8	819	3552
October 12	10:53	27	20.92	138.40	11.5	826	3576
October 19	11:35	27	22.29	150.20	12.4	821	3558
November 24	16:10	28	2.21	230.46	4.7	819	3552
Total						**11496**	**49806**

RGB-thermal images from each of the 14 flights, highlighting the presence of varying illumination and weather conditions within RT-Trees. Table A.1 reports the flight start time, flight duration, sun elevation, sun azimuth, and air temperature at the time of each drone flight, along with details on the number of images captured and processed for training. The variation in weather and lighting conditions caused due to different sun positions once again highlights the challenge of RT-Trees. Similarly, Fig. A.1 shows the image brightness (L) averaged across all pixels in LAB colour space, distinguished by flight date. Images from flights later in the day (e.g., October 7) are typically darker than those taken closer to noon due to a lower sun position. The exception to this is November 24, where the significant presence of white snow cover is inflating the average brightness level.

The total imaged area of one of the flights (August 30) was geographically split, with around 25% reserved for the test set, 5% for the validation set, and the rest for training, as illustrated in Fig. A.2. Images from the same training area from all other flights were included in the training set, while images from the testing and validation

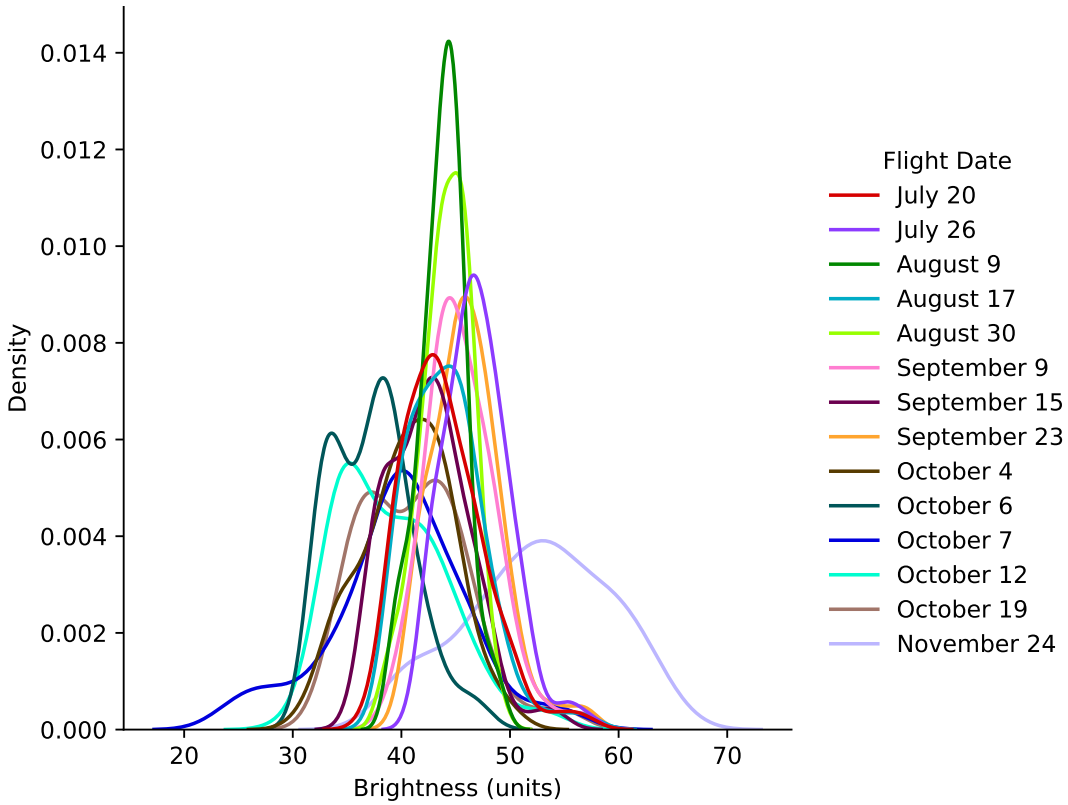


Figure A.1: **Kernel Density Estimation Plot for Average Brightness** for images in each flight date, mapped first to LAB color space.

areas were discarded, leading to roughly 70% of the total captured images being used for training. To allow for higher batch sizes during training, each training area image (RGB and thermal) was first split into six 500×500 patches as demonstrated in Fig. A.3, retaining the high degree of overlap between neighbouring images. On the other hand, only the central 500×500 patch was considered for each image in the evaluation sets, and we sampled every third image in the capture sequence from these sets to eliminate overlap, resulting in 10 patches for validation and 63 for testing. Only non-overlapping images from a single flight date are included in the testing and validation sets to ensure that each tree only appears in one image in those sets so that detection performance is not overestimated. On the other hand, the inclusion of overlapping imagery from multiple dates helps promote diversity in the training set. This explains the seemingly large disparity in the number of training and validation/testing images.

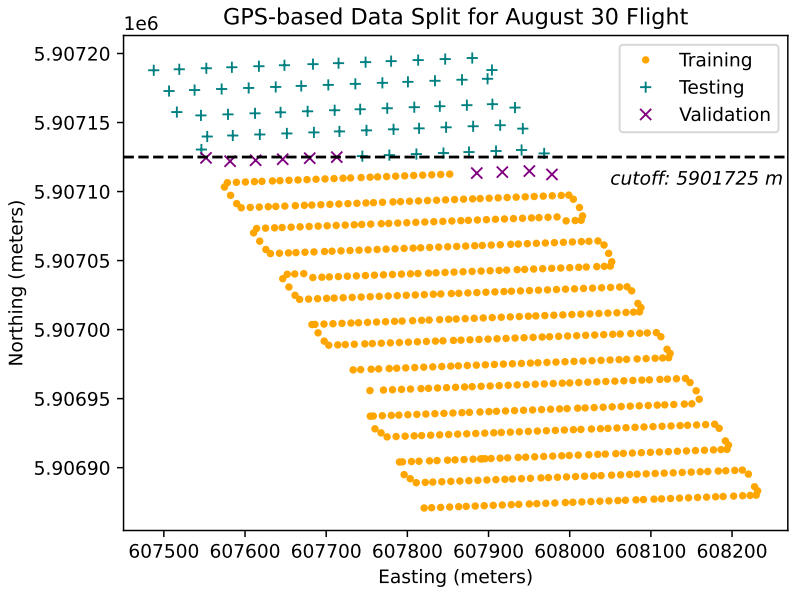


Figure A.2: **GPS-based Data Split**. Each point represents the location where a drone image was taken. The assigned cutoff line separates the testing area from the training (and validation) area.

The proposed method is entirely self-supervised and does not require ground truth annotations for training. Therefore, only image pairs in the testing and validation sets were annotated with bounding boxes in a two-step manner for evaluation purposes. First, visible tree crowns were delineated from the RGB image through careful inspection. Then, the corresponding registered thermal image was used to identify shadowed tree crowns that had been missed in the first step – these new boxes were marked as *"difficult"*. In total, 447 out of the 3611 tree crowns in the testing set were marked as difficult. In general, the difficult boxes were fewer in number (see Fig. A.4) and of a smaller area (see Fig. A.5) than non-difficult boxes corresponding to visible trees. This is because younger, smaller trees are more likely to be hidden in the shadows of their taller neighbours. Although the annotated bounding boxes were primarily square (1:1 linear relation in Fig. A.6), a considerable number of rectangular boxes are present in the testing set due to the presence of different species with non-circular crowns, partial tree crowns at the edges of images, and overlapping canopies in the densely forested region, another challenge posed by the proposed dataset.

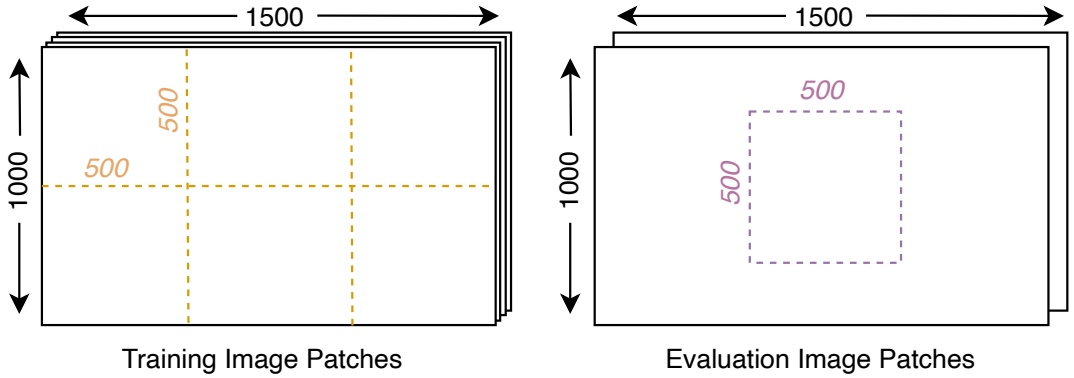


Figure A.3: **Splitting Training and Evaluation Images** into patches. All training images are evenly split into six patches, whereas every third evaluation image (testing & validation sets) is centre-cropped.

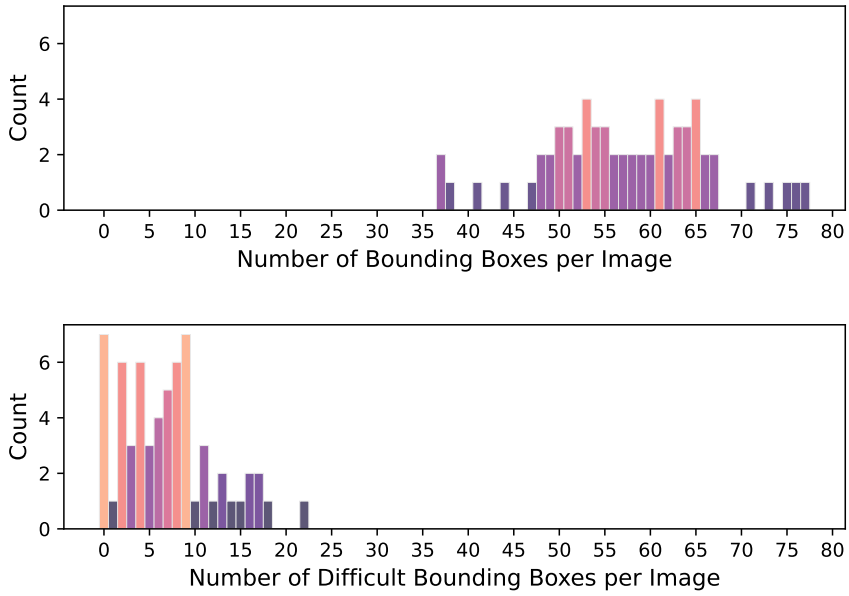


Figure A.4: **Distribution of Bounding Boxes per Image** for all boxes (top) and only difficult boxes (bottom) in the testing set.

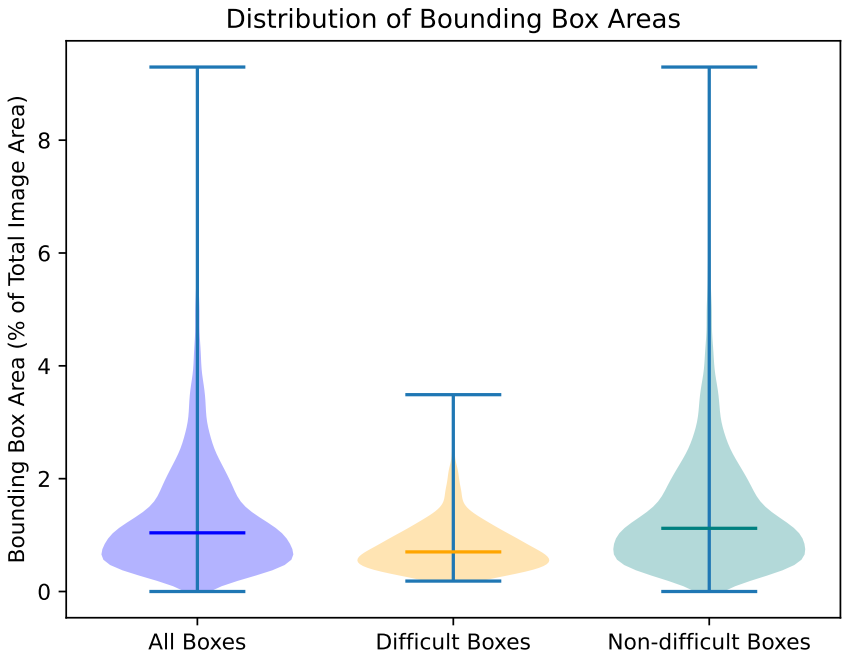


Figure A.5: **Distribution of Bounding Boxes Areas** for all boxes, difficult boxes only, and non-difficult boxes (i.e., visible in RGB image) in the testing set.

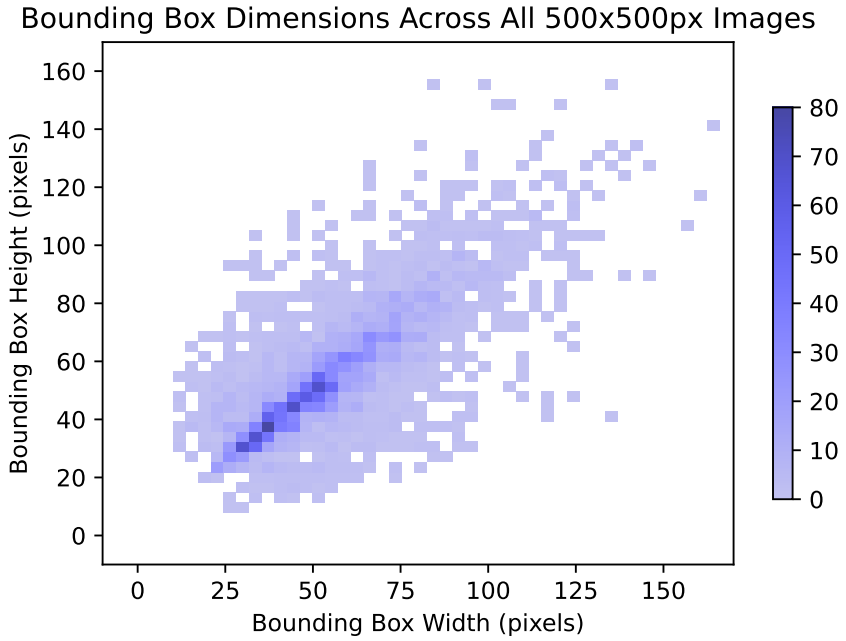


Figure A.6: **Distribution of Bounding Boxes Dimensions** for all boxes in the testing set.

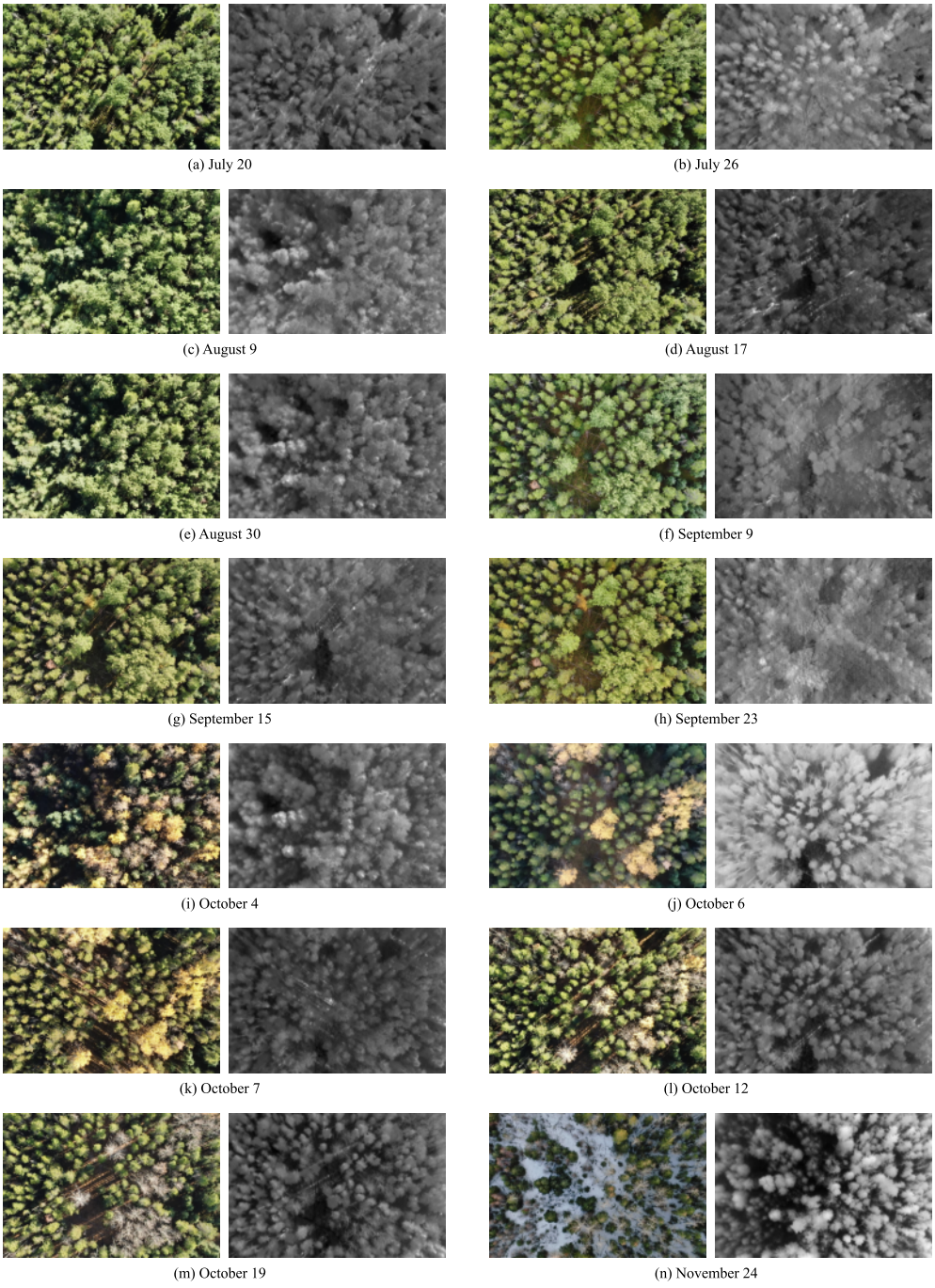


Figure A.7: **Example of Drone-collected Image Pairs** for each flight date after performing co-registration.

Appendix B: ShadowSense Extended Ablation Study

Comprehensive ablations were performed to support the selection of hyperparameters in the ShadowSense method proposed in Chapter 3. The quantitative results obtained with different values within the proposed method on the validation set are presented in Table B.1. This configuration of the ShadowSense model represents the best-performing combination of hyperparameters. Each configuration's performance was assessed independently while setting all others to their best-performing values. Different distributions for the FPN alignment scales β and fusion scaling weight η were tested. In both cases, a descending order of values (largest to smallest FPN feature map) with medium variance was found to perform the best in terms of all three metrics considered. In general, assigning higher scales to the smaller feature maps (ascending order) performed worse than when using a descending order of scales, since smaller feature maps are of a lower resolution. The classic image masking procedure used to generate binary masks relies on watershed segmentation [119, 124]. The performance of this mask generation process depends on the initial choice of thresholds for the defined BG/FG markers, and (20,100) was found to be the best choice among the alternatives. These alternatives had either less or more differences between the thresholds, which made it more difficult for the algorithm to correctly determine the marker for pixels with intensities in between. Thermal weight λ is used during inference to compute the weighted average of the FPN feature maps from the RGB and adapted thermal branches. Assigning a lower weight to thermal features than RGB features did not

help enhance the identification of shadowed trees, though foreground performance is somewhat satisfactory. As the thermal weightage was increased, shadowed tree performance increased and eventually leveled off. Using a weight of 5 was found to be the best choice, which further increases the performance on foreground trees. NMS values around 0.10 were tested, according to the recommended default value in the baseline RGB-only detection model considered [136], and 0.15 yielded the best performance.

Table B.1: **Extended Ablation Study** for different hyperparameter settings in the proposed method based on AP50 and AP100 metrics, trained without annotation on the RT-trees training set. Results on the RT-Trees validation set are reported. While changing one hyperparameter, all others are set to their best-performing values as described in the implementation details for the proposed ShadowSense configuration (also emboldened here).

Hyperparameter	**Value(s)**	**All Trees**		**Shadowed Trees**
		% AP50 (↑)	**% AR100** (↑)	**% Identified** (↑)
FPN Alignment Scales β In order of largest (lowest) to smallest (highest) FPN level.	`[1.0, 1.0, 1.0, 1.0, 1.0]` Identical Weighting	54.02	24.73	19.20
	`[1.0, 1.0, 0.75, 0.5, 0.25]` Descending w/ Low Variance	54.20	24.97	15.20
	`[1.0, 1.0, 0.5, 0.05, 0.01]`Descending w/ Med. Variance	**55.48**	**25.86**	**20.80**
	`[1.0, 0.5, 0.1, 0.01, 0.005]` Descending w/ High Variance	53.20	24.55	20.00
	`[0.25, 0.5, 0.75, 1.0, 1.0]` Ascending w/ Low Variance	50.96	24.70	12.80
	`[0.01, 0.05, 0.5, 1.0, 1.0]` Ascending w/ Med. Variance	50.21	24.61	13.60
	`[0.005, 0.01, 0.1, 0.5, 1.0]` Ascending w/ High Variance	48.49	23.41	8.00
Fusion Scaling Weights η In order of largest (lowest) to smallest (highest) FPN level.	`[1.0, 1.0, 1.0, 1.0, 1.0]` Identical Weighting	54.62	24.28	19.60
	`[1.0, 1.0, 0.8, 0.6, 0.4]` Descending w/ Low Variance	54.84	24.88	**20.80**
	`[1.0, 1.0, 0.5, 0.2, 0.2]` Descending w/ Med. Variance	**55.48**	**25.86**	**20.80**
	`[1.0, 0.5, 0.2, 0.05, 0.01]` Descending w/ High Variance	51.71	23.00	20.00
	`[0.4, 0.6, 0.8, 1.0, 1.0]` Ascending w/ Low Variance	52.09	22.75	17.60
	`[0.2, 0.2, 0.5, 1.0, 1.0]` Ascending w/ Med. Variance	50.86	22.26	16.00
	`[0.01, 0.05, 0.2, 0.5, 1.0]` Ascending w/ High Variance	48.81	21.07	10.40
Intensity Thresholds: Min. & max. in mask generation.	`[30, 75]` Low number of initially unmarked intensities	49.86	21.74	8.00
	`[20,100]` Medium number of initially unmarked intensities	**55.48**	**25.86**	**20.80**
	`[10,125]` High number of initially unmarked intensities	51.67	22.23	18.40
Thermal Weight λ_T**:** Used in weighted fusion during inference.	`0.5` Higher weighting for RGB features	50.57	22.65	13.80
	`1.0` Identical weighting for RGB and thermal features	50.31	22.03	16.00
	`2.5` |	52.61	23.08	18.40
	`5.0` ↓	**55.48**	**25.86**	**20.80**
	`7.5` Higher weighting for thermal features	52.01	22.86	**20.80**
Non-max Suppression: NMS threshold used in training and inference.	`0.05` Less overlap in filtered predictions	49.15	22.25	17.20
	`0.10` |	50.50	23.30	16.00
	`0.15` ↓	**55.48**	**25.86**	**20.80**
	`0.20` More overlap in filtered predictions	50.60	23.97	17.20