UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

OBJECT DETECTION IN DUAL-BAND INFRARED

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JOHN R. JUNGER III
Norman, Oklahoma
2023

OBJECT DETECTION IN DUAL-BAND INFRARED

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Joseph P. Havlicek, Chair

Dr. Choon Y. Tang

Dr. Ronald D. Barnes

Dr. Tomasz Przebinda

Acknowledgements

I would like to express my appreciation for the support of my wife, Whitney, and our three children: Jack, Tess, and Ronin. Their love and enduring support have given me the ability to complete this work.

I would also like to acknowledge the profound patience and invaluable advice of my adviser, Prof. Joseph Havlicek. Without his assistance in research, writing, and editing, this work would not have been possible. I greatly appreciate the effort he has put into educating me and setting me on this rewarding research path.

I owe a significant debt of gratitude to Dr. Nicholas Mould for sharing his abiding love of computer vision and DSP algorithms. He provided me with good advice and encouragement at every stage of the PhD process.

Dr. Chuong Nguyen should also be recognized for his invaluable assistance in developing my research skills and implementing code.

I would like to thank the members of the Intelligent Transportation Systems Lab at OU during my time there. They also contributed to my academic development and sense of community.

I extend my gratitude to every member of the OSSM community; they have been lifelong friends and supporters of mine. Within that community, I would especially like to thank John "Jack" Gleason. His advice to me at such a young age has had a profound impact on my academic career and interests.

I am grateful to Thordur Runolffson, who had a profound impact on my understanding of state estimation, system identification, nonlinear systems, and

control theory. His presence and intellect will be missed.

Lastly, I would like to express my gratitude to Dr. Scott Austin, who was of utmost importance in my academic life while I was at Texas A&M. I regret that he is no longer with us.

Table of Contents

List of Tables

List of Figures

Abstract

OBJECT DETECTION IN DUAL-BAND INFRARED

John R. Junger III, Ph.D.
The University of Oklahoma, 2023

Supervisor: Joseph P. Havlicek

Dual-Band Infrared (DBIR) offers the advantage of combining Mid-Wave Infrared (MWIR) and Long-Wave Infrared (LWIR) within a single field-of-view (FoV). This provides additional information for each spectral band. DBIR camera systems find applications in both military and civilian contexts. This work introduces a novel labeled DBIR dataset that includes civilian vehicles, aircraft, birds, and people. The dataset is designed for utilization in object detection and tracking algorithms. It comprises 233 objects with tracks spanning up to 1,300 frames, encompassing images in both MW and LW.

This research reviews pertinent literature related to object detection, object detection in the infrared spectrum, and data fusion. Two sets of experiments were conducted using this DBIR dataset: Motion Detection and CNN-based object detection. For motion detection, a parallel implementation of the Visual Background Extractor (ViBe) was developed, employing Connected-Components analysis to generate bounding boxes. To assess these bounding boxes, Intersection-over-Union (IoU) calculations were performed. The results demonstrate that DBIR enhances the IoU of bounding boxes in 6.11% of cases

within sequences where the camera's field of view remains stationary. A size analysis reveals ViBe's effectiveness in detecting small and dim objects within this dataset.

A subsequent experiment employed You Only Look Once (YOLO) versions 4 and 7 to conduct inference on this dataset, following image preprocessing. The inference models were trained using visible spectrum MS COCO data. The findings confirm that YOLOv4/7 effectively detect objects within the infrared spectrum in this dataset. An assessment of these CNNs' performance relative to the size of the detected object highlights the significance of object size in detection capabilities. Notably, DBIR substantially enhances detection capabilities in both YOLOv4 and YOLOv7; however, in the latter case, the number of False Positive detections increases. Consequently, while DBIR improves the recall of YOLOv4/7, the introduction of DBIR information reduces the precision of YOLOv7.

This study also demonstrates the complementary nature of ViBe and YOLO in their detection capabilities based on object size in this data set. Though this is known prior art, an approach using these two approaches in a hybridized configuration is discussed. ViBe excels in detecting small, distant objects, while YOLO excels in detecting larger, closer objects. The research underscores that DBIR offers multiple advantages over MW or LW alone in modern computer vision algorithms, warranting further research investment.

Chapter 1

Introduction

Integrating new sensors into field deployable systems is of key practical interest in engineering. Many of these systems are intended to operate in real-time with limited computing resources. This dissertation aims to characterize one such sensor, namely a Dual-Band Infrared (DBIR) image detector. To do so central contributions to the field will be the introduction of a new labeled DBIR data set for use in object detection and tracking and strong evidence to support the benefits of DBIR sensors for object detection with modern object detection algorithms. A key design limitation is that the object detection algorithms must be capable of real-time performance. A second design limit will be that the detection algorithm must be usable off-the-shelf if data is not available for modification.

Object detection in images is an ongoing field of active research. Object detection algorithms with the highest performance are many times slower than real- time, this issue is discussed at length in the literature review. A second substantial limitation is that some types of modern detection algorithms are significantly better at detecting medium and large size objects but struggle at detecting small or distant objects. Since the focus of this research is to make a usable and deployable system we will look at other classes of detectors which are better at detecting small and distant objects. The research in this work

will show that DBIR substantially improves performance of the first class of detectors especially as the size of the object decreases. Further it will show that DBIR substantially improves the detection of smaller objects in a second detection algorithm and that these detection algorithms are complimentary when it comes to sensitivity to size. The last, and perhaps most important, thing demonstrated here is a new data set that made the previous two contributions possible. This new dataset will also make possible further research into the extremely important class of image-based DBIR sensors. A substantial collection of DBIR video sequences is presented, accompanied by a comprehensive description of the target objects featured within these sequences. DBIR data finds utility across various domains of application, ranging from self-driving cars to remote sensing satellites. While literature pertaining to the production of DBIR sensors, lenses, and supplementary equipment continues to grow steadily, the discourse concerning the information derived from these sensors has dwindled. This dissertation aims to reverse this trend. With preliminary evidence highlighting some of the benefits of DBIR data, along with the introduction of a new dataset for evaluating these advantages, it seeks to rekindle research into the benefits of enhanced detection and tracking capabilities facilitated by the incorporation of additional spectral information.

1.0.1 Problem Statement

Around 20 years ago, Dual-Band Infrared (DBIR) was a topic of considerable research interest [54, 55]. During that period, the prevailing notion was that DBIR offered numerous advantages, particularly in enhancing battlefield perception by augmenting detection capabilities under low-light conditions. The

subsequent decade was characterized by the central idea that DBIR could extend detection capabilities within visually challenging environments, such as dusty and smoky conditions. For a comprehensive list of applications, please refer to the literature review in Chapter 2.

In 2012, two distinct groups embarked on efforts to collect datasets: one, a collaboration between American military research laboratories and NATO allied researchers [190]; the second, the laboratory at the University of Oklahoma, where the research reported in this document was conducted. Since then, the realm of computer vision has witnessed profound transformations, yet the literature related to DBIR has not kept pace with these significant changes. While the body of literature pertaining to the design and manufacturing of DBIR systems maintained its previous momentum, the research presented here aims to rectify this research gap. It does so by evaluating modern motion and Convolutional Neural Network (CNN) based detection systems using two-color data, an endeavor novel to the DBIR body of literature.

Contemporary computer vision detection algorithms heavily rely on the curation of extensive datasets [48,109]. Few of these extensive datasets cater to infrared [71,175], and none encompass DBIR. To train and validate these CNNs, these datasets require object location labeling or "ground truth." Although the dataset presented here remains relatively small compared to the dataset sizes necessary for substantial CNN retraining, it provides compelling evidence that DBIR data confers performance enhancements in more modern algorithms than those featured in the most recent DBIR evaluation publication. The collection and annotation of vast datasets across diverse contexts entail substantial costs. Demonstrating a performance advantage using a preliminary DBIR dataset for

contemporary computer vision algorithms helps justify the expense associated with dataset collection. This dissertation illustrates the existence of a preliminary dataset and underscores that, even without extensive CNN retraining, these datasets yield advantages with multiple contemporary object detection algorithms.

Chapter 2

Literature Review

In this chapter, the following points will be addressed:

1. **Definition of "Detection":** The term "detection" will be defined and contextualized within existing literature.

2. **Significance of Detections:** The importance of studying detections within sensor systems will be established.

3. **Contextualize Object Formation from Detection:** How raw sensor detections become understood as objects.

4. **Applications of DBIR:** An examination of the applications of Dual-Band Infrared (DBIR) will be provided, considering its additional complexity in modern sensor systems.

5. **Metrics Terminology Clarification:** The terms False Alarm (FA), Clutter, Confidence, and Signal-to-Noise Ratio, Average Precision (AP), and mean-Average Percision (mAP) will be defined.

6. **Sensor Fusion Importance:** The significance of fusing information from multiple camera sensor modalities will be established.

7. **Challenges in Fusion:** The challenges in fusing information across camera sensing modalities will be discussed, along with the benefits provided by multi-spectral sensors.

8. **IR Systems as Detectors:** The utilization of Infrared (IR) systems as detectors will be explained.

9. **Detection via Images:** The utilization of images for detection purposes will be elaborated.

10. **Detection via Video:** The utilization of video, motion detection in particular, will be discussed.

11. **Measurement of Detection Capabilities:** The methods for measuring detection capabilities will be covered.

Studying the detection capabilities of Dual-Band Infrared (DBIR) poses a significant challenge due to extensive terminological overlap with systems not intended for IR digital image creation. For instance, there are numerous papers within the SPIE (International Society for Optics and Photonics) related to dual-band radar [51,70,126,134,185], as well as examples of dual-band/dual-frequency lasers [76, 97, 125, 135, 184]. Researchers must engage in substantial filtering efforts to navigate this terminology. Additionally, the majority of DBIR papers pertain to the manufacturing of Focal Plane Arrays (FPAs), optical filtering systems, or the introduction of new dual-band camera systems to the market, rather than their capabilities.

Another complicating factor is that the term "Dual-Band" can refer to any combination of Visible-spectrum (VIS), Electro-Optical (EO) [7, 15,

100, 146, 158], Near Infrared (NIR) [1, 26, 33, 34, 37], Far Infrared (FIR) [11], Long-Wave Infrared (LWIR) [129, 183], Mid-Wave Infrared (MWIR) [129, 183], Short-Wave Infrared (SWIR) [24, 65, 67, 94, 177], and UV [25, 79]; this includes DB-LWIR/LWIR [57]. The references in the prior sentence are small samples of papers and should not be understood as complete lists. For this study we are specifically interested in MWIR/LWIR Dual-Band Infrared (DBIR). That there are so many papers about production and manufacturing of DBIR components is strong evidence of the belief and/or knowledge that DBIR systems have a significant future or present non-published application. Hudson and Hudson discuss the constraining impact of the US military classification system on the open literature and their own need to omit certain aspects [69]. There is a high likelihood that many applications of DBIR are not in the open literature as the dominant literature is US Army Research Labs [56] and NATO research [190].

Among the numerous conference papers and journal articles on Dual-Band Infrared (DBIR), three emerge as particularly significant for this dissertation. The first pivotal paper is titled "Application of dual-band infrared focal plane arrays to tactical and strategic military problems" by Arnold Goldberg, Theodore Fischer, and Zenon Derzko [55]. In this paper, the authors delve into a limited dataset of DBIR images captured in the MWIR/LWIR spectra. The images encompass military vehicles, individuals, buried landmines, and missile exhaust plumes.

The second central paper originates from the same research group at the Army Research Lab (ARL) and is titled "Analysis of Dual-Band Infrared Imagery from the Multidomain Smart Sensor Field Test" [54]. This paper applies specific analysis techniques to the dataset acquired in the preceding "Applica-

tion..." paper [54]. In [54], the authors employ a Multi-Layer Perceptron (MLP) to classify detections as either background or targets. It's noted that the detector's performance is highly sub-optimal, and the "chips," or sub-images, representing targets, are frequently neither centered nor complete representations. The authors also highlight the substantial likelihood of overfitting due to the limitations of their data and training. For each testing category, the authors utilized fewer than 300 target sub-images and under 2,000 clutter "chips." The authors discuss employing Principle Component Analysis (PCA) to expedite the training process. This technique involves raster vectorization of the sub-image, creating high-dimensional points. Subsequently, the covariance matrix of these points is generated and undergoes an eigen-value decomposition. The eigen-vectors associated with the highest eigen-values correspond to the most information-rich features within the dataset. The Multi-Layer Perceptron (MLP) is subsequently trained on the eigen-vector training set, reducing the volume of images that require inclusion in the training process. "Analysis..." also encompasses several other significant applications, including Automatic Target Recognition (ATR) and perceptual enhancement through image fusion. The MLP aspects discussed in this paper were reported in [24], where the contributors presented the detector and MLP techniques. "Application of dualband infrared imagery in automatic target detection" [28] features a figure showcasing the eigen-targets, which could be of interest to readers. As the authors of [28] articulate, ATR encompasses "preprocessing, detection, segmentation, feature extraction, classification, prioritization, tracking, and aimpoint selection." However, the paper predominantly focuses on the detector and classification aspects of ATR. The significance of tracking as a consumer of de-

tections will be discussed later, and it's worth noting that Automatic Target Recognition (ATR) closely approximates the Joint Directors of Laboratories (JDL) sensor fusion model discussed later.

In [55] the image fusion process, information from the Mid-Wave (MW) and Infrared (IR) channels is merged into a common Red-Green-Blue (RGB) visual spectrum display, as opposed to grayscale images from each channel being presented separately. This technique aims to provide enhanced vision to soldiers and tank crews by presenting information from multiple sensor systems within a single view on the display. Notably, this type of image fusion is designed for human consumption and is evaluated by analyzing the ease or difficulty of target detection in the enhanced versus unenhanced images [55]. It's important to note that the image fusion techniques discussed in [55] are not assessed for the detector introduced in the present work. However, they present a valuable avenue for future research, potentially enhancing the performance of non-human detection algorithms in consideration of perceptual enhancements from image fusion.

Several distinctions exist between the study conducted in this dissertation and Application [55] and Analysis [54]. In this dissertation, more modern Convolutional Neural Network (CNN) are employed for infrared imagery, as opposed to MLP used in [54]. This study validates the use of CNN detection without requiring extensive training or transfer learning. Moreover, this study incorporates the potentially superior motion detection algorithm ViBE [12]. Notably, variations in target categories exist between this dissertation and [54] featuring humans, while this study includes civilian vehicles and aircraft, vice [54] that focuses on military vehicles, landmines, and missile

exhaust. In a similar manner, the authors of [28], who are also researchers at Army Research Lab (ARL), describe the process through which the MLP for clutter rejection was generated for [55]. The paper faces the same challenges related to detector localization as reported in [55].

The last significant paper dealing directly with the topic of this dissertation is "A standard data set for performance analysis of advanced IR image processing techniques" authored by A. Robert Weiß et al. [190]. The significance of this paper stems from its introduction of the North Atlantic Treaty Organization (NATO) Data Set-140, the only other "published" dataset of images available. However, due to its controlled nature, accessing this dataset presents challenges. I have initiated contact with the authors of this paper, requesting access to the dataset or, at the very least, some information about it. As of the time of writing, I have yet to receive a response from them. In Reference [190], the paper includes two published targets: people and military vehicles. The discussed sample applications encompass Dynamic Range Compression and Super-resolution. Super-resolution involves the endeavor to enhance pixel density/size in an image by extrapolating information between recorded pixels, typically achieved through the utilization of a Generative Adversarial Network (GAN) [68].

The applications of Single-Band IR are too numerous to comprehensively list in this context. However, a more concise compilation of applications specifically utilizing DBIR can be provided:

1. Strategic Systems for Early Warning of ICBM Launches [69]

2. Standoff Detection of Poison Gas [69]

3. Aids for the Precision Delivery of Weaponry [69]
4. Passive Infrared Guidance of Air-to-Air Missiles [69]
5. ATR [54]
6. Detecting People [55]
7. Land Mine Detection [55]
8. Strategic Targets (Missiles) [55]
9. Detection of Tactical Vehicles [55]
10. Inspection of Bridge Decks [81]
11. Perception enhancement [54]
12. Remote Sensing (RS) [193]
13. Super-Resolution (SR) [190]
14. Small or Dim Target Detection [141]

2.1 Detection

In "The theory of signal detectability" by Peterson et al., the definition of "detection" is provided as the identification of a "signal plus noise," as opposed to the presence of "noise" without a signal [145]. The theory presented in [145] builds upon the groundwork of communication theory as introduced by Claude Shannon in "The Mathematical Theory of Communication" [161]. Both [145] and [161] outline the process of message generation by the receiver, as illustrated in Fig. 2.1. The determination of a "signal plus noise" indicates that

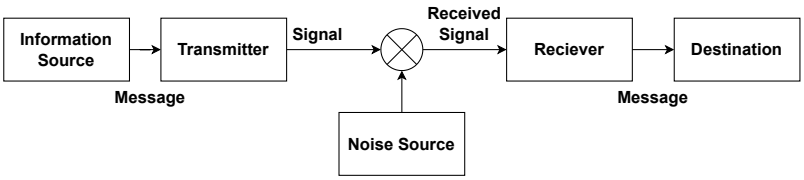


Figure 2.1: Shannon Communication Channel

there is an attempt at communication within a communication channel, or that the presence or location of an "object" has been sensed in a sensor system. In this way a "sensor system" acts like a receiver in a communications channel and a "detection" is the determination that there is a signal present. Within the context of this work, the distinction between the act of detection and the reporting of detection holds less significance. Generally, a signal that is determined to not be a detection is not conveyed at the receiver's level, ensuring that the recipient of the message is not burdened with filtering out genuinely negative responses.

The fundamental difference between Shannon's communication theory and the theory of detection lies in the purposeful engineering of the information source and transmission medium to facilitate efficient information transmission. In communication channels the attributes of the transmitter are extensively characterized through engineering procedures, allowing the design of a channel that enhances the probability of successful signal reception [161]. This entails substantial engineering efforts to guarantee accurate information transmission and recording. For instance, Error Correction Codes (ECC) incorporate extra bits to minimize errors and maximize error-free storage on hard-drive disks [120].

However, various communication scenarios arise where designing to sup-

port the receiver becomes infeasible or undesired. In military applications, adversaries employ techniques like frequency hopping or visual camouflage to complicate signal detection and make it more challenging [163]. Further challenging the detection problem is the open nature of the transmission medium exposing vulnerabilities to manipulation by signal hijacking, or denial via jamming [101, 163]. In different contexts, receivers might utilize "signals of opportunity" to formulate a received message [40]. These signals of opportunity are often suboptimal for detection purposes. In these situations the only part of the communication system that can be efficiently engineered is the receiver, i.e. sensor.

Signals such as reflected light, unintentionally emitted acoustic energy, and radio emissions can be harnessed to develop detections, paralleling how biological systems utilize sense organs. Sensor systems of this kind are classified as "passive" since they do not emit a signal and await its return for detection [60]. The data set presented here in Chapter 3 falls into this category. A survey of the sales literature for MWIR and LWIR shows that they typically operate in passive operations with reflected or emitted light. Short-Wave IR on the other hand can be operated passively but often includes an illuminator.

In contrast, active sensors emit energy and measure the reflected energy. Common examples of active sensors in modern systems encompass Sonar, Radar, and LIDAR. Active sensors offer two primary advantages over passive sensors. Firstly, they can ascertain range through time-of-flight calculations, and secondly, sophisticated active sensors can generate distinct signals that stand out from background noise. In many contexts emitting energy is undesirable because it will assist an adversary in localizing the source of the

emission. Other non-military restrictions on emissions are present, for example emissions of Radio Frequency Energy are restricted by governmental agencies beyond certain energy levels to facilitate communications in those spectra as well as safety to humans [95].

Despite grounding the definition of "detection" in first principles, we still encounter challenges in the practical study of detection, largely stemming from systems engineering practices developed to overcome limitations to communications channels engineering. I define "detection modality" as the physical model of the sensing mechanism, the transmission media, and the signal processing of the received signal. Then complexities arise when single modality sensing systems are used for object detection and tracking. One approach to improve performance of sensor systems is to add sensors using different modalities and combining the information from multiple sensors. This process is called "data fusion" [8] or sometimes "data association" [102].

Given Shannon's model of the communication channel as in Figure 2.1 the "noise source" implies that sensors are imperfect. Significant research has gone into how to improve the reliability of these channels [121]. When the engineer has control of the signal generation the signal can be sent multiple times or in a unique way to lower the noise floor by isolating the transmission medium from noise sources. Many practical applications, especially those using sensors outside of controlled environments, can not take advantage of engineering the transmission channel and must accept the transmission media as it is. Assuming that noise is uncorrelated a short duration detection may be a considered a false alarm, where persistence indicates the detection is more likely to be a real, signal [8, 41]. In this way time plays a significant role in the

detection methodology and time-series analysis is central in feature definition. The topic of time series analysis as it applies to improving signal detectability is too deep and rich to cover sufficiently here, but there are many books, articles, and journals dedicated to the exploration of this important topic; for example see [171] but almost all texts covering sonar or radar signal processing will address these issues. In many sensing systems a natural byproduct of determining if an input is "signal plus noise" or "noise" is how closely it matches some exemplary signal [171]. The closeness of this match or the strength of the signal relative to the noise level can be used to determine whether a signal is present. The presence of the signal plus noise may not be sufficient to meet all of the systems needs; in some cases localization information is also necessary. Typically this localization is relative to the sensor and positional information of the sensor is needed to comprehend the signal generator's position. Another often necessary piece of information is a quantitative measurement of the relationship between he strength of the signal to the strength of the noise. This is quantified by Signal-to-Noise Ratio (SNR), a confidence model [75, 87, 205], or a clutter model [200]. Clutter modeling has its background in radar, especially ocean borne [77, 118, 200] but also in other sonar and other military sensor systems [50, 202].

A note on terminology: There is a close relationship between the terms "measurement" and the term "detection." In Figure 2.1 the "Receiver" is a device that measures the state of the communication channel. In one sense this is a "measurement" and a "detection" is what happens after some processing occurs that determines that there is actually a signal present. Data fusion often operates within a system-of-systems architecture and what is considered

a "detection" within the limited domain of a single sensor may be considered a "measurement" at a higher level of data refinement [8]. Detections are often fused into tracks as part of the data refinement processes of a single sensor [8, 41], and then fused again into tracks between distributed or sensors with dissimilar sensing modalities [8, 41]. In this way the detections at one layer of the system become measurements at the next layer. This is all to say that the language is muddled because the processes used in association and refinement may repeat themselves at additional layers [8, 41].

An individual detection is sufficient to determine the presence of an object and to temporarily establish its location. However, it's often useful to ascertain whether a new detection corresponds to an object previously detected or represents an entirely new object. Additionally, understanding how the object is moving relative to the sensors holds significance, and this activity is known as "tracking" [171]. The practice of tracking objects has a rich history in both military and commercial applications [8, 171]. Tracking is a deep and rich subject, but to select one approach from many, Multi-Object Tracker (MOT) [18,188] is an excellent approach associating detections in environments without significant engineering controls. An MOT typically operates in three stages: a data-association stage, a track update stage, and a track creation stage. When a new detection is received, the MOT checks if the detection can be associated with an existing track. If so, the track is updated with the new information. If the detection cannot be associated with any existing track, a new track is typically created with that detection as its initial data point. Most trackers are probabilistic, often following Bayesian principles, and they provide information about the mean and covariance of the tracked

object [102]. Two popular association techniques are called Gaussian Nearest Neighbor (GNN) and Joint Probablistic Data Association Filter (JPDAF) [9]. In GNN an appropriate metric between the detection and the track estimate is used to determine if the detection is close enough to the track to be associated with it citeGNN. JPDAF updates all tracks with the weighted probability of association, a special track initiation step is required [9].

The space within which the objects are being tracked depends on the sensors involved. Many MOT operate in a spatial coordinate system such as Cartesian, polar, or spherical coordinates [171]. This is an important factor because the data association procedures at L1 often use spatial information, that is if a detection is located at the same point in space and has the same kinematic attributes as a track being maintained then that detection has a high probability of being part of that track [8]. The most likely space for the data presented in this work is "pixel space" the location within the two-dimensional image of the object. If the tracker is Bayesian the likelihood that a detection belongs to the track can be evaluated by measuring the Gaussian probability that the detection is from the track's current mean and covariance [102]. "Tracking" and "data fusion" are often treated as the same problem in keeping track of the location of detections of many sensors over time can be thought of as fusing information from each of those sensors into a track [8]. Without going into too much detail here, in the Bayesian approach the estimated location of the "signal source" is updated to the time of a sensor detection, then using an association method similar to the one above it can be determined if the detection belongs to a given track [86, 102]. That track can then be updated with the information [102]. These techniques can be applied to multiple instances of

the same sensor such as in a "sensor network" or in dissimilar sensors [39, 128]. Increasing the number of sensing modalities can provide several benefits. It can increase the opportunity for detection and, if the noise characteristics are independent between the sensing modalities, be used to reduce the number of false alarms [39]. A third benefit is the generation of "appearance" models generated by the detections signature in multiple sensing modalities which might be used to improve the association procedure [107].

2.1.1 False Positive Suppression

A false positive is a detection not associated with a real signal source or the desired kind of signal source. I will follow [8] and consider "clutter, clutter points, random clutter, false alarms, false detections, false measurements" and "false positives" to be interchangeable. A distinction can be made here between transient or uncorrelated FP and a "persistent" FP. A persistent FP is a source of FP that sustains in times. A persistent FP in our context may be an unlabeled stationary object that was not labeled in the data set because it is not of interest. This will generate FP throughout the sequence without being associatable to ground truth objects.

Another common association approach is the use a Multiple Hypothesis Tracker (MHT). An MHT is used in either a post-processing mode or a windowed history approach. MHT treats the association problem as an optimization problem on the historical detection data set [152]. In "An algorithm for tracking multiple targets" by Donald Read, the goal is to try to associate every possible "measurement" with every possible target. Each possible combination of tracks and measurements are then presented as a hypothesis. Through op-

timization detections are associated with the highest probability tracks [152]. Another hypothesis, that the detection is a FP, called the "null hypothesis" is also considered in this approach. In MHT a track exists through the history of the window and the more detections that can associated with a possible track increases the likelihood that the track is real and that the detections associated to the track are True Positive (TP) [152].

However, deploying MHT techniques face a significant challenge: many of them involve NP-Hard optimization problems that are often too computationally expensive for real-time operations. This computational burden limits the practical application of these methods in real-world scenarios [152].

Rather than using MHT other types of tracking engines can be employed.Some commonly used ones include Particle-Filter (PF) [72,139], Kalman Filter (KF) [18, 188], as well as variations of KF such as Unscented Kalman Filter (UKF) [80] and the Extended Kalman Filter (EKF) [80]. To keep the explanation concise, we will assume that a suitable tracking algorithm is available to perform the task of data association and track creation. Given as suitable tracking engine for an MOT an approach that I have used to account for the amount of information the track has been updated with and the recency of those updates. If a track has enough "recent" information, i.e. recent detection that track can be reported. If a track fails to have enough information content, then the track is considered "clutter" and is not reported. This method is similar to the MHT with the exception that it does not optimized the detection to track association by considering all possible considerations but can maintain track association using less computationally expensive techniques like GNN or JPDAF.

If the information content of the track drops too low the track is considered "stale" or "dead." The removal of "stale" tracks is a common part of the track maintenance of MOTs [8]. This concept can be likened to the notion that tracks need to be "fed" with new data to remain active; tracks that are not updated "starve" and eventually die out. In [8] there is a thorough treatment of track maintenance and FP suppression.

In many scenarios, clutter detections - unwanted or false detection - tend to be uncorrelated in time. By applying this form of time filtering, the reports of uncorrelated clutter detections can be eliminated. This filtering approach helps improve the accuracy of the tracking results by removing spurious or unrelated clutter reports from the tracked objects. This section discussed techniques to remove clutter-decrease the number of False Positives-in the next we will look at how to increase the number of True Positives. As Section 2.8 describes the metrics the main goals of sensor system design and system of sensors engineering is to increase the number of TPs and decrease the number of FPs. One thing that should be considered is that the removal of False Positives via tracking almost universally also removes true some TPs as well. This creates a trade-off between the competing goals of reducing FPs while maximizing TPs. Detections contributing to tracks that have not acquired enough information to be reportable will not be reported, and when the track becomes reportable some number of detections will have been lost to time [8, 41, 152]. Another potential issue is that persistent clutter sources will not be effectively filtered using these techniques [8, 41].

If DBIR data increases the number of TP detections of objects it will increase the performance of trackers of this sort especially at track initiation.

Then employing these techniques central to data fusion and tracking will amplify the utility of these types of sensors.

A widely used method in the CNN computer vision community for reducing false positives is Non-Maximum Suppression (NMS). NMS is effective in decreasing the number of highly similar detections produced by CNNs. Implementations of popular models such as YOLOv4 and RCNN typically include NMS capabilities in their distributions [20, 153].

The NMS algorithm follows this procedure: for a given collections image detections, select the one with the highest class confidence. Calculate the IoU of this detection with all other detections and remove those surpassing a certain threshold. Next, choose the detection with the highest IoU among those not yet selected or removed. Repeat this process until no more detections remain unselected or unremoved. The algorithm systematically eliminates detections too similar to others, prioritizing those with higher class confidence.

2.2 Systems of Sensors

Four of the main limitations of individual sensors are modality, dimensionality, false alarm rate, and field-of-view [8, 41]. Overcoming these limitations in theory increases the number of True Positive detections, a central goal. Adding sensors with different modalities can be used for clutter rejection of sensors in different modalities [41]. Generally speaking sensors with uncorrelated detection modalities have uncorrelated clutter [8]. A target may have strong emission signatures in more than one sensing modality and some combination of sensing modalities may comprise a more unique fingerprint, or as mentioned before "appearance," for a given target [8]. The number of modes used in

detection can also be built into the clutter model and tracks with detections from multiple modes can be given a higher confidence and thus be reported sooner and with more confidence. Bar-Shalom offers a detailed conversation of geometric models in [8, 41]. He also describes in detail how to construct measurement models and multi-sensor fusion [8].

Certain sensors may have lower-dimensional data or narrow fields of view but possess advantageous qualities such as lower latency. In theory, it's beneficial to utilize sensors with wider Field-of-View (FoV) to cue sensors with narrower FoV to detections [41]. This strategy can enhance detection frequency and minimize undesirable attributes like high latency. When sensors are not collocated, data fusion techniques can be employed to increase the dimensionality of tracks through stereo sensors [8, 41]. Data fusion to support these considerations is a highly active area of research, particularly in applications such as self-driving or assisted-driving automobiles [14, 44, 84, 144], which may also involve the integration of additional sensors like LIDAR and Inertial Measurement Units (IMUs) [166, 187].

More closely related to the work at hand the fusion of visible spectrum data with infrared (IR) data has garnered substantial research attention over the past decade and a half [41, 61, 137, 148, 194, 206].

To provide context and clarify the positioning of the sensor under investigation within the broader sensor architecture, let's consider its features and integration. The Dual-Band Infrared (DBIR) data and the camera responsible for capturing it belong to the category of "passive" sensors, with limited fields-of-view (FoV) [41]. To leverage the benefits of this system, it would likely be mounted with mechanical alignment to either a stationary mount or a Pan-

Tilt-Unit (PTU) [41]. The PTU enables a control system to direct the field of view towards specific search areas, and components within the PTU provide spatial measurements of the camera's pointing direction. In many military applications, high-precision wheel encoders are used on high-end PTUs like the Night Hawk Position by PVP Advanced Systems [147], while less expensive PTUs may employ stepper-motor positioning like the FLIR PTU-D48E [176]. Optical encoder based PTUs can acheive angular resolutions in the tens-of-thousandths of degrees [147] where stepper motor based resolutions are in the hundredths of degrees [176]. In my experience higher precision PTUs translate to higher precision measurements when other sources of error are controlled like mechanical engineering tolerances and backlash. If the DBIR data is intended to be used in conjunction with non-collocated sensors, a method would be required to determine the orientation of the other sensor relative to the DBIR camera [41].

In some cases, integrated camera systems are available on the market that mount both EO and IR sensors within a single PTU camera system [147]. Later in our discussion, we will explore the challenges associated with fusing data from camera with disparate modalities. However, an advantage of the DBIR camera system is that its different sensing modalities, specifically MWIR and LWIR, share a Common Optical Axis (COA), and the correlation between the pixels in the two modalities is trivial to establish. This alignment simplifies the process of associating corresponding features between the two modalities for fusion purposes.

A model for sensor fusion, known as the JDL model [8, 41, 170], was developed to provide a framework for understanding and discussing data fu-

sion processes. The JDL model was initially crafted by the Directors of the US Army, Navy, and Air Force Laboratories, with the aim of simplifying the language used to describe sensor systems and their data fusion activities.

The JDL process model structures the data fusion process into five levels all of which are quoted from the original text:

1. Level 0 (L0) — Sub-Object Data Assessment: estimation and prediction of signal/object observable states on the basis of pixel/signal level data association and characterization;

2. Level 1 (L1) — Object Assessment: estimation and prediction of entity states on the basis of observation-to-track association, continuous state estimation (e.g. kinematics) and discrete state estimation (e.g. target type and ID);

3. Level 2 (L2) — Situation Assessment: estimation and prediction of relations among entities, to include force structure and cross force relations, communications and perceptual influences, physical context, etc.;

4. Level 3 (L3) — Impact Assessment: estimation and prediction of effects on situations of planned or estimated/predicted actions by the participants; to include interactions between action plans of multiple players (e.g. assessing susceptibilities and vulnerabilities to estimated/predicted threat actions given one's own planned actions);

5. Level 4 (L4) — Process Refinement (an element of Resource Management): adaptive data acquisition and processing to support mission objectives

Using the JDL model to structure our thinking, here we are most interested in L0, fusing information in the pixel space, or L1, fusing information after detection has been made in the MWIR and LWIR pixel spaces separately. Multiple spectra in the same physical space ameliorate many L0 and L1 fusion challenges. For images generated from separate sources, these are both challenging activities. The literature explaining these challenges is discussed in Section 2.3.1 and Section 2.3.2. It is worth looking at the literature to understand what challenges can be avoided by combining sensing modalities in the same Focal-Plane Array (FPA). In practice, these various layers of the JDL are often conducted by separate systems after L1 [8, 41, 170]. Usually, sensors are built as separate sensors, and engineers combine information from these into a centralized or distributed processing system [8, 41, 170]. These are combined through the fusion process to make multi-modal tracks and to understand that objects tracked by different systems are the same [8, 41, 170]. The distinction between L0 and L1 deals mainly with "objectness." More information can be useful in determining whether an some signals are an object or clutter, they can also be useful in forming a model for the appearance of a model. In the Section 2.3 some of the complication of fusion at L0 are discussed. The fundamental nature of detection in the JDL model give weight to the importance of characterizing the detection capabilities of new sensors on the novel data sets they generate, like the one presented here. Data can be fused at either L0 or L1 and there can be feedback processes between these layers [41]. How tightly integrated the sensors are in the data fusion process is a deep and rich area for study, an excellent source for further reading is [41].

2.3 Information Fusion

In the previous section we discussed the JDL model for data fusion. In this section, we will look at the current techniques for fusing information at L0 and L1. In DBIR systems, the need to fuse information is still present, and approaches to fuse it can be seen in some of the earliest and most important literature on the topic such as the "Application" [55] and the "Analysis" papers [54] which discuss how to fuse the separate channels of a DBIR system into a single visual display.

From 1980 to 2022, the IEEE published approximately 2000 articles on image fusion. Multi-modal image fusion is still a very challenging and research-worthy topic. A second central problem in multiple camera systems is to understand the content of images taken from multiple perspectives, i.e., lacking a COA.

The previous sections discussion of the two main parts of the JDL model apply to multi-view camera systems as well. The information can be fused at the L0 or L1 levels [41]. At L0 information is fused within the pixel plane to generate a detection, and at L1 the information is fused after the "detection" has been generated. Even though the sensor systems examined in this document do not require these techniques because systems with a COA do not require this type of fusion it is worth examining the techniques used to achieve this level of information fusion to understand the added complexity of systems without COA.

Two main approaches exist to fuse information at L0, image registration and image disparity mapping [22, 52, 83, 122, 140, 167, 203]. At L1 after the

detection has been formed traditional sensor fusion techniques can be utilized as found in [8]. Image registration determines the scale and positional transformation of one image to the other to overlay the information between two or more images [22, 122]. Disparity mapping is similar to image registration but tries to build a map of portions of one image to another [140, 167]. Traditional sensor fusion techniques were discussed in Section 2.2.

In the next two sections, I will argue that DBIR offers significant advantages in L0 information fusion over systems with two separated LW and MW camera systems. Stereo techniques at L0, specifically image registration and disparity mapping, are computationally expensive and pose significant hurdles when applied to DBIR data [173]. However, L1 stereo techniques can be applied without modification to systems with separate MW and LW detectors [41]. However, while "stereo-like" techniques can be applied to perform sensor fusion in COA, it should be noted that a COA configuration is inherently ill-suited for retrieving depth information because the "baseline" or "pupillary" distance is zero. For a detailed description of deriving depth from the geometry and resolution of depth the reader is referred to [52].

2.3.1 Image Registration

The aim of Image registration is to align images of the same scene using the information contained in the images [59]. In the literature, research in image registration has a heavy focus on medical imaging [22, 52, 122, 203]. Medical imaging registration has added complications due to 3-dimensional scanning techniques used by medical imagers. Remote sensing, particularly the analysis of satellite images, is another focus of image registration technique

research [110, 204]. The output of image registration is most often the transformation of one image to another such that the pixels of the transformed image correspond most closely to the pixels of the other image. For video sensors image registration can be used for video stabilization, where sequential frames are aligned and cropped to smooth the effects of motion in the camera. Image registration is also often used between sensors to correlate information between two sensor systems. These systems do not need to be the same sensing modalities [130, 204]. In [130] the authors, all from the Air Force Research Lab (AFRL), make the case for multi-layered and multi-modal information fusion and evaluate four different techniques for image registration: Lucas-Kanade optical flow, Ohio State University Correlation method, Robust Data Alignment (RDA), and Scale Invariant Feature Transform (SIFT). The Ohio State University Correlation method referred to above is a pyramid based approach that does template matching between the reference image and the image to be transformed [131]. Perhaps, the most well known algorithms for image registration are Scale Invariant Feature Transform (SIFT) [116] and Speeded Up Robust Features (SURF) [13]. SIFT and SURF are both based on "keypoint matching" which has a long history in image processing [62]. In [62] a method to generate features- or keypoints- using edge and corner detectors is described. The contribution of [13] and [116] are to make these features rotation and scale invariant. To get good performance from the keypoint matching algorithms the algorithms require many keypoints in both images and then a Nonlinear Least-Squares (NLS) optimization is performed to minimize the error between the associated keypoints distances [115]. NLS tends to be a computationally expensive operation [115]. Another significant difficulty talked about in more

depth in Subsection 2.3.2 is the difference in textural perception in different camera sensing modalities. Unlike image fusion between sensors of the same modality, edges and corners are textural features and often do not correspond in different spectra.

Another commonly applied image registration is the maximization of the mutual information between images. This technique is common in medical image registrations [22, 122, 203] and remote sensing [4, 110, 164]. The techniques typically involve geometric translation of the "floating image" and then comparing the mutual information with the "reference image." Mutual information is described in Claude Shannon's *A Mathematical Theory of Communication* [161], but the concept is named by Robert Frano [88]. Mutual information is the relationship between two random variables. It quantifies how well the probability distribution of one random variable can be known from observing another [4, 110, 122, 164]. The probabilities being described are almost exclusively in the pixel intensity space [4, 110, 122, 164], while [22] discusses the problems with pixel-intensity-only registration and discusses three papers that make attempts to add spatial information [22]. In [22] many of the mutual information based technique limitations are discussed namely: ignorance of spatial information, can not take geometry into account, sensitivity to noise, high cost of computation especially high-dimensional medical images, and contour misalignment if the feature space is reduced prior to apply mutual information techniques. The actual calculation of Mutual Information or the related Kullbeck-Leibler (KL) divergence are interesting in establishing how much information difference there is in the intensity information between spectra but the amount of mutual information or divergence between the spectra

is itself not particularly useful for the work at hand. Establishing the fact that the probability functions of the pixel intensity are dissimilar can be established quickly upon visual inspection of Fig. 4.11 and Fig. 4.12.

2.3.2 Disparity and Correspondence Mapping

Disparity mapping is the process of generating three-dimensional scene geometry from stereo camera systems. In most cases the camera systems have known intrinsic camera parameters which are the internal camera optical and geometric characteristics [64]. The extrinsic parameters such as the pointing vectors of the camera axis and their spatial relationship to each other must also be known to a high degree of accuracy [64]. Then the images from the two separate sensors are used to estimate the spatial locations of object with the mutual fields of view of the sensors [173]. What stereo disparity matching does is provide is a fully-dimensioned representation of the mutual fields of view of the cameras. In this way passive sensors are able to estimate depth via geometry at individual pixel locations. The visual disparity, and similarly the ability to resolve depth, is relative to the extrinsic parameters of the camera system. The most salient parameters are the distance between the cameras, and the distance to the observed object. The remoteness of the camera system to the feature, in as an example Remote Sensing (RS), where the distance between camera systems co-mounted on satellites are a small fraction of the distance to the observed objects on the surface of the earth [181] minimizes this problem.

Disparity mapping typically requires determining the corresponding sections or patches between the images. In this way disparity mapping algorithms are also stereo correspondence maps and pixels can be related between fields

of views and similarly images [173].

Common difficulties found in Disparity Matching algorithms are noise, textureless regions, depth discontinuities, and occlusions [173]. The authors of [173] point out that correspondence matching "is an ill-posed problem with inherent ambiguities." This often leads to an incomplete correspondence mapping with large gaps of uncorresponded pixels. From personal experience working on these problems the difficulties listed above are exascerbated when the modalities of the camera systems are not the same because different camera modalities observe differing textures given light transmission properties and emissivity. This observation plays out in practice fusing IR to VIS which is an active field of research [83, 140, 167], but textural differences between MWIR and LWIR can be plainly seen in, e.g. in Figs. 4.22, 4.23, and 4.24.

2.4 Infrared Literature

The literature about pure IR detectors is significant and pre-dates the second World War when IR detection systems were starting to become viable. Many of the early papers on IR Detection are about L0 of the JDL model above. That is, they discuss how the presence of IR photons are detected at the transducer level. However, several articles related to detection and search are available from that time. In [74], Jamieson described three circuits used in IR detectors. He goes on to describe the act of detection as "Optimum detection then requires the testing of a hypothesis that the sample function was drawn from a population of signals and noises, against the hypothesis that the sample function was drawn from a population of noises alone [74]." This definition perfectly mirrors Peterson via Shannon's definitions in [145, 161]. Jamieson discusses the operation of

"Matched Filters." Jamieson describes building an optimal filter under the conditions that the noise is additive, Gaussian, white, and band-limited. He goes on to describe computing the cross-correlation between the input and the "expected signal," something that we would now call a "feature." Then a threshold would be applied to determine if the input has enough similarity to the feature to be considered a detection or not. Jamieson also discusses background rejection in this paper using time-multiplexed or non-colocated IR sensors. A third circuit he describes uses both the background rejection circuit and a second memory circuit to understand the location of the detections to provide an output of when new signals are detected and display to the user. The concepts parallel much of the discussion below about using DBIR as a sensor.

In [69], Hudson and Hudson give a brief summary of the state of military applications in the IR bands of the spectrum. This paper gives thorough explanation of the use of NIR as well as Midwave and Longwave IR. The paper also discusses the performance increase in detecting Inter-Continental Ballistic Missile launches in IR outside of the NIR bands. A very thorough explanation of IR transmission in Earth's atmosphere is provided with an excellent visualization of O_2, H_2O, and CO_2 absorption notches in the spectrum. These effects can be seen in the differential transmissibility in our data set between the MWIR and LWIR bands given the greater depth of field in the LWIR compared to MWIR. The shift from NIR to FIR also allowed the seeker heads in IR sensitive missles to track the plume of jet exhaust [74], allowing a missile to track an aircraft from various angles of attack. Given the absorption bands of NIR this posed a difficulty because of cloud based and solar glinting causing

high levels of false detection in NIR [74].

In [99] using IR to detect aircraft to create a collision avoidance system, however somewhat novelly the authors suggest adding an IR beacon to aircraft to assist in detection when aircraft is not oriented in a way to easily detect the exhaust plume. This is a case of engineering the signal as well as the detector. In [133] the authors use a series of image processing techniques to find connected-component objects in IR images. The objects that were detected were M-48 tanks. It should be noted that some spectra of IR tend to blur edges making this a particularly difficult task. Connected components labeling will be used in our evaluation of DBIR in Section 2.6.

In 1983, [136] started discussing "small target" detection. The context of [136] was space-borne radiometric measurement of aircraft against a cluttered IR background, namely sea-water from space. This is important to note because small-target has become an emphasis of IR detection in the modern era as small targets are specifically difficult for CNNs to detect by classification [30, 49, 119, 193, 198]. "Infrared small target" represents its own body of research literature. The application of CNNs to IR images has altered the detection approaches, but CNNs have some small size limitation and engineers will always be trying to push the boundaries of detection size. By the 1980s the miniturization of IR detectors had gotten to the point that supporting space based IR detection could be achieved for the detection of ICBMs [127]. In 1989, [172] proposed a single cell IR photodiode blind-spot warning system for drivers which would involve the detection of automobiles. The earliest example of using a Neural Net to detect objects in IR images is in 1989 [155]. While that literature will be reviewed separately in Section 2.7.4 it is worth noting the date here for

reference. In [98] the authors apply motion detection to FLIR images using a multi-resolution pyramid approach and correlation energies. In the IEEE 1988 and 1989 seems to be the time when researchers started looking into more complex feature-spaces for detection with [35] and [66].

2.5 DBIR and IR Super-Resolution

Given the relatively low pixel count of DBIR, and until recently of MWIR and LWIR camera systems, a topic of focus in the IR/DBIR communities has been on IR Image Super-Resolution [68, 113, 162, 189]. Super-Resolution is the process of increasing the number of pixels in an image, or to increase the resolution of an image from lower pixel-count to higher pixel-counts. Traditional methods for this are interpolation, generally by polynomial splines but over the last 15 years success has been found in the use of Generative Adversarial Networks [68, 113, 162, 189]. Significantly this is also an approach to detecting small and distant objects.

The FPA used to collect the data set described in Chapter 3 has two pixel-level deficiencies. The first is that some pixels in the MWIR were damaged prior to acquiring the sensor, I will call these "stuck pixels" because their value does not change across all sequences. The second phenomena is that the calibration of the temperature regulation process was sensitive to miscalibration, I call these "frozen pixels." Fixing the stuck MW pixels and frozen pixels might be a research worthy application of GAN to this dataset. Super resolution research could also be accomplished by training a GAN on this data set after sub-sampling.

2.6 Detection from Video: Motion Detection

Moving object detection holds central importance in image and video processing [12, 112, 124]. Moving object detection has been proposed to improve another class of algorithms we will discuss later, namely CNNs [112]. The region proposal/detection algorithm for Region Convolution Neural Network (R-CNN) is discussed in Section 2.7.1. Probably the simplest form of motion detection involves the simple differencing of subsequent frames. If frames I_t and I_{t-1} are subsequent frames in a video sequence then the pixel-wise difference $\Delta_t = I_t - I_{t-1}$ is a way to detect the changes between two frames [199]. More complicated frame differencing algorithms take into account more information, like an increased number of frames, utilizing more intricate finite difference methods.

Another perspective is that the information extracted from frames I_{t-1} ... I_{t-k}, where k represents a range of past frames, can be seen as a model for the background. How the information is extracted from these past frames indicates the type of model. Certain models are called "parametric" in that they try to create a parameterized background model in a statistical sense. That is to say that the moments of the statistical distribution of the background model are stored as parameters. For example, a probabilistic model might characterize the background as a Gaussian function and store the mean μ and co-variance σ. A common approach models the background as multiple Gaussians and is known as a Gaussian Mixture Model (GMM) [201]. One of the drawbacks of parametric models like the GMM is that they require complex pixel-wise linear or non-linear regressions [168, 169, 201], which can be be ill-posed. Given that the background model will likely be updated with each incoming frame,

parametric fitting models encounter challenges when operating in real-time due to the necessity of linear regression. This can easily number in the millions of pixels with modern camera systems.

Other parametric models which are popular in the literature update their parametric models "on-line" by updating their statistical model with only the latest information [124]. The Σ-Δ filter is one of the latest and highly performative examples of this approach [112].

Up to this point, the discussion has centered around detecting motion itself rather than the detection of objects. In the frame differencing discussion we referred to Δ_t as the difference map. As a matter of convenience some arbitrary distinction must be made to declare if a pixel location within the difference map is "in motion." For example $M_t = |I_t - I_{t-1}| > \epsilon$ or $M_t = (I_t - I_{t-1})^2 > \epsilon$ where ϵ represents a predefined threshold. M_t is a binary map of pixels that are in motion and not in motion. For the GMM and Σ-Δ models discussed above the map M_t determines the probability that the incoming pixel belongs to the background, i.e., not in motion or if that pixel is in motion.

Another approach to the background modeling problem is a "nonparametric" approach [12, 112]. Non-parametric methods store a collection of samples and subsequently compare the incoming frame to that collection of samples avoiding the need for expensive pixel-wise regressions. The algorithm for Visual Background Extractor (ViBe), a non-parametric model of importance, is central to Section 4.1.3 so a thorough explanation of the algorithm is presented here.

Up to this point, I have discussed motion detection or background sub-

traction but not object formation. There are various methods for object formation, with the most straight-forward and likely most common being grouping of detected "moving" or "foreground" pixel location into objects. It is common to miss individual pixel movement or foreground pixels and to have small aberrant foreground detections. Classic image processing approaches are used to clean the image; typically morphological erode and dilate operations [52]. Then a process known as Connected Components Analysis (CCA) is done to find connected regions in the motion plane M [112]. For background reading on basic image processing techniques see [52]. These connected components can then be treated as detections by either reporting the pixel locations as a "segmentation map" or calculating the "bounding box" for the pixels. Within the pixel space, these representations effectively localize the detected object.

2.6.1 ViBe

In [12] Olivier Barnich and Marc Van Droogenbroeck describe a non-parametric visual background extractor called ViBe. The authors of ViBe discuss the domain of literature for background subtraction and the utility of Σ-Δ in embedded processors because of its lack of need for floating point calculation. The approach taken in ViBe is that the background model $M(x) = \{b_1, ..., b_N\}$ stores N samples per pixel location taken from the previous frames when the pixel locations are labeled "background." When an incoming pixel-value $I(x)$ at location x is compared to each value being stored at $M(x)$. If $|I(x) - b_i| < R$ for $i \in \{1...N\}$ a counter is incremented. If that counter is greater than some third parameter $\#_{min}$ then the pixel location x is labeled background otherwise labeled foreground. This relatively simple approach has proven to be highly

effective. There is a fourth parameter that controls the rate at which the background model learns incoming data designated as ϕ. ViBe also has a feature which updates the neighboring pixel locations. This feature is especially useful to reduce false positives caused by camera sway or other harmonic phenomena within the camera FoV.

A sketch of a the parallelized version of ViBe that I used in my research is available in Appendix A. In experimentation neighborhood learning was not found to have a great effect on the performance of ViBe and was omitted for speed and convenience. Given that the visual effect is a slight blurring of the sample space another option would be to add a slight Gaussian blur to the sample image and have every pixel learn from every other pixel. The Gaussian kernel would need to be tight. GPUs are adept at filtering images quickly and applying this blur could be done very quickly given that the parallelized version of ViBe is being used in the GPU.

2.7 Deep Learning and Convolutional Neural Networks

Over the past decade, research in object detection has been largely dominated by deep learning techniques, particularly CNN. Deep Learning (DL) is the use of Artificial Neural Network (ANN)s with more layers than traditionally used in MLP. A CNN is a Deep Neural Network that includes convolutional filter layers inside the networks. These convolution filters are typically learned on a labeled dataset and evaluted on a training dataset. The convolutional layers represent learned features that are applied to the image and neural connections between the layers learn how the combinations of these feature output to establish classification. There are two types of CNN based object detectors one-stage

like YOLO [151], SSD [114], and RetinaNet [150] or two-stage like R-CNN [53] and Faster R-CNN [153]. R-CNN networks apply a region proposal network and a classifier network sequentially. Most detection algorithms of this type classify proposed regions on how well they match classes within their trained dataset [149,153]. Then the last layer outputs the class and likelihood that the detection is part of that class. This information about the confidence that the detection belongs to the class can be used to evaluate whether the detection is more likely a TP with high confidence, or a FP with low confidence [153]. You Only Look Once (YOLO) divides the images and classes as subimages. Then YOLO evaluates each section of the input image and evaluates if it could be part of a class. The classes are then aggregated on the output layers to generate the detection. In R-CNN often thousands of regions, which are effectively sub-images are passed through the classifier [53]. YOLO on the other hand only passes the whole image through once resulting in a reduced execution time.

2.7.1 Datasets and Network Performance

There are several standard datasets for training and performance evaluation of CNN object detectors. A survey of the data sets is listed in Table 2.1. Three of these data sets stand out as the most important for benchmarking CNN the first is Microsoft Common Objects in COntext (MS COCO) which includes 328,000 images and 1.5 million object instances with segmentation masks some with multiple objects per image [109]. The second is the PASCAL Visual Object Classes the challenge for which was operational 2005-2012 [48]. The Pascal VOC server is still open for submissions and as of the time of writing the highest scoring leader is a YOLO variant and the second is an R-CNN variant.

The third important dataset is called IMAGENET [43, 91, 156]. IMAGENET has 14,197,122 images with the last organized challenge being 2017 [43].

2.7.2 YOLO

Joseph Redmon et al. published 'You Only Look Once: Unified, Real-Time Object Detection" in May, 2016 [149]. The innovative feature of YOLO is that it unified the detection and classification network. This leads to real-time performance relative to the frame rate of normal 24-30 frames per second (FPS) cameras. In [149], YOLO is compared to R-CNN, Fast, Faster-R-CNN, Deep MultiBox (DMB) and OverFeat [159] on Connected Components Analysis (PASCAL VOC) 2007. DMB and OverFeat are faster than real-time while Fast R-CNN, Faster R-CNN, and R-CNN are slower than real time. Of the faster than real-time detectors YOLO outperforms the others by more than twice in terms of Mean Average Precision (mAP). Of the slower than real time YOLO is outperformed by some versions of Faster R-CNN by approximately 0.1 mAP with the network Faster R-CNN with VGG-16 [153] achieving only 7 FPS. In [149] it is reported that YOLOv1 can achieve a 63.4 mAP at 45 FPS. For two-stage detectors mAP and speed are inversely correlated and the computation time is often not accounted for in many of the larger competitions with the highest mAP. YOLOv2, a.k.a. YOLO9000, out performs Faster R-CNN with Resnet as the classifier on PASCAL VOC and MS COCO at 67 FPS [150]. In 2018 Redmon and Farhadi released YOLOv3 [151] which is described as an incremental improvement and that he had "phoned it in for a year." The paper reports better performance of YOLOv3 compared to RetinaNet [151]. After this publication Joseph Redmon stepped away from developing YOLO.

Table 2.1: Modern Computer Vision Data Sets

Microsoft COCO: Common Objects in Context [109]
The Pascal Visual Object Classes Challenge: A Retrospective [48]
Long-Term Visual Object Tracking Benchmark (TLP) [138]
Visual Tracker Benchmark (VTB) [195]
Object tracking benchmark [196]
The visual object tracking vot2015 challenge results [89]
The visual object tracking vot2014 challenge results [90]
Free FLIR data set: https://www.flir.com/oem/adas/adas-dataset-form/s [73]
The Amsterdam Library of Ordinary Videos (ALOV300++) [165]
Temple Color 128 (TC128) [108]
NUS/BUAA People and Rigid Objects Dataset (NUS)/(BUAA) [103, 104]
AMCOM: FLIR data set
DARPA Video Verification of Identity (VIVID)
Short Term Single Object STSO, integrated into VOT 2015...
Multi-Object Tracking Benchmark [42, 96, 132].
Military Sensing Information Analysis Center (SENSIAC)
DAVIS: Densely Annotated VIdeo Segmentation [82]
NII Okutama-Action: An Aerial View Video Dataset [10]
LITIV PTZ Tracking [31]
LITIV Thermal-Visible Registration [17]
LITIV-Single Object Tracking dataset [23]
Ess *et. al.* Multi-Person Tracking [47]
TUB MOCAT [19]
Mouse Embryo Tracking Database [36]
Berkeley Motion Segmentation (BMS-26) [174]
Freiburg-Berkeley Motion Segmentation (FBMS-59) [142]
Korea Advanced Institute of Science and Technology (KAIST)EO/IR [71]
FREE Teledyne FLIR Thermal Dataset for Algorithm Training [175]

YOLOv4-v8

You Only Look Once version 4 (YOLOv4) was introduce in April of 2020 it represented a significant improvement over You Only Look Once version 3 (YOLOv3) on the MS COCO dataset and outperformed similar CNNs per given use of computation time [21]. On this data set it achieves 43% Average Precision (AP) over 60 FPS. Th highest performing in terms of AP in [21] was EfficientDet but at just under 50% AP but that was slightly over 10 FPS and the performance of EfficientDet falls below that of YOLOv4 at just over 40 FPS.

After YOLOv3 Joseph Redmon decided that he would no longer develop YOLO because he did not want to contribute to the military applications announced in a tweet found here:
https://twitter.com/pjreddie/status/1230524770350817280.

The three years between 2020 and 2023 become somewhat anarchic with YOLOv5 being announced on June 25th of 2020 by Glen Jocher [78]. Mr. Jocher promised a paper in Dec 2026 in the same repository, as of the time of writing that paper is not available [78]. YOLOv4 and You Only Look Once version 7 (YOLOv7) (2022) were created by the same team. YOLOv6 results were published by a completely separate team from Meituan Inc [105]. The authors compared the performance of YOLOv6 with YOLOv5 and YOLOv7 [105, 186] concluding that YOLOv6 outperforms v5 and v7.

2.7.3 Vanilla, Fast, and Faster RCNN

In [53], Girshick et al. introduce their method called "R-CNN: regions with CNN features." The central idea is that any region proposal algorithm can be

used and then the regional sub-samples can be fed to a CNN for classification. Regions that do not exceed a certain threshold for belonging to a classification are rejected. The region proposal algorithms suggested by Girshick et al. are Support Vector Machine (SVM) [27], objectness [3], selective-search [182], multi-scale combinatorial grouping [6], and sliding-windows. The central problem is that objects in images can be of any scale and at any pixel location within an image and for a modern mega-pixel image this constitutes tens of millions of possible sub-images. In effect these region proposal algorithms pass thousands of sub-images into the classification network. In Girshick et al. their region proopsal method provides "around 2000 bottom-up region proposals." The region proposal algorithm selected was an SVM. They trained on PASCAL VOC 2007 dataset, fine-tunecd on PASCAL VOC 2012 and tested against PASCAL VOC 2010. Their results show that they are competitive against other contemporary region proposal based networks. Faster R-CNN is an attempt to lower the number of sub-images by introducing Region Proposal Networks (RPN) [153]. Per Ren et al., Fast R-CNN only made the classifier faster and didn't address the region proposal issue. The RPN is trained end-to-end with the classifier network via stochastic gradient descent on the MS COCO dataset. The RPN shares several convolutional layers with the classifier network when using R-CNN. They conduct several experiments where they use a selective search SVM to train the classifier network on PASCAL VOC then train the classifier network. They report achieving approximately 17 FPS [153].

2.7.4 Applying CNNs to Infrared

Using a CNN model trained on visible spectrum images such as PASCAL VOC or MS COCO and then using them to detect/classify infrared images is common practice [73, 106]. In [73], Jiang et al. use this method and evaluate them on the FLIR Dataset. In [106], Li et al. retrain YOLOv5 with some modifications to the network architecture specifically on the KAIST data set. Fortunately they report the results of their retrained network in comparison with YOLOv3 and YOLOv4 not retrained. Li et al. show that it performs 19.1 AP% better than the non-retrained networks. Unfortunately, they don't compare it to unretrained YOLOv5 so it is not obvious how much of that contribution is due to the different detector/classifier network and how much is due to the retraining on the dataset against which it is going to be tested. In [111] Liu et al. compare Faster R-CNN trained on LWIR data to Faster R-CNN trained on EO data. The results show that the improvement gained by retraining is marginal [111]. The value has to be inferred from the Precision/Recall curves to a few percent mAP. Other papers show a large improvement in thermal detection via retraining such as [92]. A significant problem with [92] is that as, I will discuss in the next section, MS COCO has relatively few small and distant targets [32, 85]. In [160], the authors compare YOLOv4 in an excellent experiment that separates the visible spectrum and IR aspects with other variables of the dataset. In [160], Shaniya et al. use YOLOv4 to do detection on an EO/NIR UAV dataset. In [29] the authors compare a YOLOv4 model trained only on MS COCO to one they trained with transfer learning on the FLIR dataset. They show about a 6.93% improvement in performance after transfer learning for people and 11.64% improvement for "cars" [29]. In the conclusion

for [29] conclusion the authors state that YOLOv4 "outperforms state-of-the-art methods without being trained on thermal images and fine-tuning leads to even better results." It should be noted that the DBIR data set that we use in our experiments has a similar perspective to some of the images in the FLIR ADAS data set [73]. In [5] the authors generate a control loop to track-over-detection with a FLIR MWIR camera using YOLOv4 as a detector without retraining or transfer learning.

2.7.5 Small Object Detection

One known drawback to CNN object detectors, besides run-time limitations, is difficulty with small objects [16, 32]. This problem is present in YOLO due to the structuring of the grid on which YOLO classify/detects. In two-stage detectors like R-CNN it is a factor of the multi-scale resolution design tradeoffs [154]. Another reason that it is difficult to detect small objects is that small objects have less information because the information is lost in the sub-pixel spaces. In [32], Chen et al. discuss "Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal" [32] as four approaches for small object detection. Super-resolution was discussed earlier in this literature review. The authors note that there are no high image count small object data sets that currently exist. This survey goes into what it might take to address the small object issue.

Small objects lack appearance information to distinguish the object from background [178]. In the data labeling process for the data set introduced in this dissertation in Chapter 3 the author had significant difficulty labeling small objects that were detectable via motion detection algorithms described

in Chapter 4. As stated in this literature review most CNNs trained on data sets labeled by people. Without employing these motion detection algorithms to the data the labelers in this work would not have been able to continue the labeling activity for very small objects. Most data sets used for training CNNs like MS COCO and Pascal VOT are still images labeled by humans [43, 48, 109]. If humans don't do a particularly good job of labeling small and distant targets other methods will be needed to label them as well. I found contextual information gained from video useful in labeling small targets in the data sets presented in this dissertation. The big datasets use still images and don't have a wealth of small objects [43, 48, 109].

Small objects can be made more classifiable by making the objects less small via the use of optical magnification. A simple control loop can be implemented for example using a non-CNN based small object detection method, for example motion detection, then zoom the camera system on that detection for classification purposes. This is effectively putting more pixels on target and increasing the amount of information for the classifier. The DBIR data set introduced here and the experiments support this approach. The sequences which we will discuss later have many long tracks where vehicles travel from great distances into the near foreground. They start as small objects and then become large objects. YOLOv4 is unable to detect/classify in the distance when the objects span only a few pixels, is able to intermittently detect when the objects are around 100 pixels in area, and detect well when the objects are much larger than that. This indicates that a control loop as described would be an automatable method for dealing with small objects. The detection range of military sensor systems can be an important selling point, for

examples see [45, 46]. In tactical/strategic operations range is among one of the important sensor characteristics [93, 143]. Detecting small objects at a distance allows those targets to be addressed earlier and eliminates the advantage of surprise.

Defining small objects will be useful for our research and as I will show there are many small objects in the data set being studied. In [32] small objects are less than 32^2 pixels, medium are not small and less than 96^2 pixels, and large objects are greater than 96^2 pixels. The authors of [32] point out a large difference in performance between these classes; the YOLO variant that they examine is YOLOv2. Other definitions of small objects include Society of Photo-Optical Instrumentation Engineers (SPIE)'s at 80 pixels in a 256 x 256 image, and MS COCO defines small to be 32^2 [117] pixels as previously mentioned. Detection by size analyses will be presented later in this work and the reader carrying forward some concept of what is a "small" object is will be beneficial to understanding that presentation.

Given the real-time performance of YOLO, the relatively good detection capabilities, and the anarchic environment around YOLO post v4, YOLOv4 and YOLOv7 were selected for the experiments in this dissertation. This decision was made to provide two examples of YOLO v4 and post v4. Variants of YOLO and R-CNN still dominate the MS COCO and PASCAL VOC challenges which makes them both good candidates. However, the speed performance of YOLO in the end makes it a more likely candidate to deploy in systems that require real time operations such as tactical field deployed sensor systems.

2.8 Metrics

With the introduction of the MS COCO and the PASCAL VOC challenges, measurement of object detection algorithm performance has converged [21, 53, 153]. The base measurement is Intersection-over-Union (IoU) which takes the intersection of the detection and the ground truth and divides it by the union of area that are covered by the detection and the ground truth. A diagram with IoU for bounding boxes is shown in Fig. 2.2. This quantity approaches one as the overlap of bounding boxes become the same. IoU approaches zero if the bounding boxes do not overlap at all.

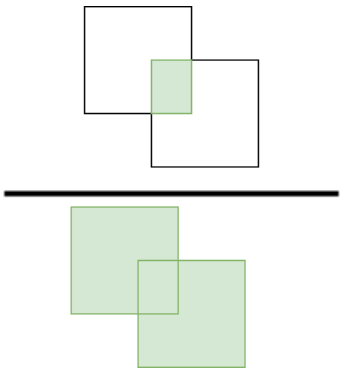


Figure 2.2: Diagram of Intersection over Union. IoU is the value of the area overlapping divided by the total area between ground truth and detection.

Most YOLO and RCNN papers compare the performance of their object detectors using Precision-Recall Curves [21, 53, 153]. The IoU is used to determine if a detection is detecting a real object in agreement with ground truth known as a True Positive (TP). If the IoU is too low the detection is determined to be a False Positive (FP). IoU establishes a relationship between the ground truth and the detections as well. If the maximum IoU between a given ground truth and all available detection is below a given threshold, the

detector is determined to have a False Negative (FN). These three values can be used to construct the important values of precision P as

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

and recall R as

$$R = \frac{TP}{TP + FN}. \tag{2.2}$$

Modern CNN-based image object detectors generate a confidence level. This confidence threshold can be varied from 0-100%. The confidence output is intended to assist the consumer of the object detection to determine if a detection is a TP or FP. For a given confidence threshold the TP, FP, and FN are calculated for all the detections with confidences above that threshold. Then the Precision P and Recall R are calculated as in (2.1) and (2.2). The (Recall, Precision) ordered pairs can then be plotted on the x-y axes. Integrating the area under the curve gives the Average Precision (AP) for the detector for a particular IoU. Most modern detectors detect many classes and the average across all classes is the mean Average Precision (mAP) [48, 109]. Better performing object detectors have ordered pairs that have higher values on the Precision-Recall plane. As a note MS COCO also supports pixel-wise segmentation which would support irregular shapes though the calculations are similar to the bounding-box based calculations suggested in Fig. 2.2. Bounding-box figures can often include background elements and thus can introduce more error than pixel-wise segmentation. The benefit of bounding-boxes is that it is a simple polygon and labeling is significantly cheaper in terms of effort.

The academic literature around CNNs for object detection has become,

in my opinion, remiss and rarely reports Average Recall (AR); for examples see [20, 149, 150]. For the sensor engineer the Average Recall, or as is called in the data fusion and tracking field Probability-of-Detection (PoD), is an important value for the data fusion process [8]. The AR for YOLOv4 was reported in [192] while retraining on MS COCO to be **0.61** with an IoU threshold of 0.50, or mAP@0.5, and **0.37** with mAP@0.75. For YOLOv7 the AR can be obtained from the github repository [191] and has a value of **0.68** over all target sizes, **0.54** for small targets, **0.73** for medium targets, and **0.84** for large targets.

Chapter 3

New DBIR Data Set

In modern computer vision, the creation and curation of large data sets is central to the performance of computer vision systems [48, 109]. The practice of generating and labeling data has developed a central importance. In this section, I provide a description of a new DBIR data set which can be used for the development of object detection, sensor fusion, and tracking algorithms. Along with the experiments run below it is the first DBIR data set to have modern computer vision algorithms applied to them in the open literature.

A total of 51 image sequences were captured. Nine sequences were captured at Brown's Campground (brwncamp) in Bishop, California. These sequences include images of civilian vehicles and people. Twenty-eight sequences were captured at the Santa Barbara Airport (SBAP) in Goleta, CA, containing birds, civilian vehicles, people, fuel trucks, and civilian aircraft. Six sequences were captured at the Von's (vons) grocery store in Bishop, California, containing civilian vehicles and people. Seven sequences were captured an intersection of Patterson Avenue and Cathedral Oaks Rd in Goleta, CA (SBPATT). Of these sequences 43 were labeled with object bounding boxes using custom built software to enable perceptual enhancement of the images for labeling.

Data labeling of these sequences with object bounding boxes was done by myself or by undergraduate students working under my supervision. All

Table 3.1: Sequences SQL Table

index	sequence_name
1	SBAP18
2	SBAP3
3	SBAP1
4	SBAP7
5	SBAP15
6	brwncamp6
7	brwncamp5
8	brwncamp2
9	SBAP11
10	SBAP19
...	...

sequences from brwncamp, SBAP, and vons were labeled. Two sequences from SBPATT were partially labeled before depleting labeling time and budget. Labeling was done in two stages: first, a sequence was labeled, and then each labeled sequence went through a quality check to verify the accuracy of the labels. A video of each labeled object was generated and carefully reviewed for accuracy and to minimize missing frames or errant bounding boxes. All of the data was added to a Sequential Query Language (SQL) relational database. Description of the SQL Tables and example SQL code is provided to the reader to facilitate the use of this data set. Sequences were given a unique identifier and linked to their names, shown in Table 3.1. There are 43 total entries in the Sequences SQL Table, some rows of the SQL tables are omitted from this dissertation to save space.

Objects were enumerated in the video sequence and were subsequently classified, or given an "object label." The classification labels for each object are

Table 3.2: Objects_Labels SQL Table

sequence_id	object	label
1	1	16
1	2	4
1	3	3
2	1	16
2	2	15
3	1	14
4	1	14
5	1	14
6	1	3
6	2	7
...	...	...

provided in Table 3.2 and correspond to an index into Table 3.1 and Table 3.3. There are a total of 233 objects in the SQL table, the first 10 are shown in Table 3.2.

The "label names" table provides human-readable labels and a mapping to the FLIR ADAS dataset classification categories. The "FLIR coco index" was added to aid in the creation of retraining data for YOLOv4 and YOLOv7. Although the retraining attempt did not yield useful results, both the labels and a complete set of YOLO-compatible annotations exist for future research. MS COCO has 80 classes which was reduced to 8 to be compatible with the FLIR dataset. The complete table is included here.

The "Spectra SQL Table" table keeps track of the definition of spectrum for future experiments. There are three total entries in this table.

The "ground truth with indexes" (gt_w_idx) table contains 68,170 entries; the first 10 are shown for brevity. The "sequence_id" column is linked to

Table 3.3: Label_Names SQL Table

index	name	flir_coco_index
1	Indeterminate	6
2	Pickup	6
3	Car	2
4	SUV	2
5	Van	2
6	Semi	6
7	Semi and Trailer	6
8	Box Truck	6
9	Pickup?	6
10	SUV and Trailer	6
11	Person	0
12	Pickup and Trailer	7
13	Motorcycle and Trailer	4
14	Airplane	4
15	Bird	
16	Fuel Truck	6

Table 3.4: Spectra SQL Table

id	spectrum_name
0	LW
1	MW
2	DB

Table 3.5: Label_Names SQL Table

pk	sequence_id	object	frame_number	x	y	width	height
1	2	1	1	161	126	29	29
2	2	1	2	161	126	29	29
3	2	1	3	161	126	29	29
4	2	1	4	161	127	29	29
5	2	1	5	161	127	29	29
6	2	1	6	161	127	29	29
7	2	1	7	161	127	29	29
8	2	1	8	161	127	29	29
9	2	1	9	161	127	29	29
10	2	1	10	161	127	29	29
...	...	...	...	...	...	...	...

the sequences table, while the "object" column links to the primary key of the object table. The "frame_number" indicates which frame is being annotated. The x and y coordinates are the upper left hand corner of the bounding box. The "width" and "height" specify the dimensions of the bounding box.

The following SQL query combines the separate tables into an easier to read view.

```
select gwi.pk,
s.sequence_name,
gwi.`object`,
gwi.frame_number,
gwi.x,
gwi.y,
gwi.width,
gwi.height,
ln2.name
from gt_w_idx gwi,
sequences s,
object_labels ol,
label_names ln2
```

```
where s.`index` = gwi.sequence_id
and ol.sequence_id = gwi.sequence_id
and ol.`object` = gwi.`object`
and ln2.`index` = ol.label
```

To evaluate performance of ViBE, YOLOv4, and YOLOv7 the detections were inserted into SQL tables and the calculation of IoU was done in the database. Table 3.6 through Table 3.28 describe the content of each labeled sequence. These tables indicate the name of the sequence, the object number, the first and last frame of the object, and the classification of the object that was labeled.

Regrettably, many of the largest thumbnails include many of the damaged pixels in the MW. The thumbnails here are taken at the largest point, but the data set has many very small data objects. While many data sets neglect labeling small objects, for example FLIR ADAS [73] which has many images where the larger objects are labeled and small objects not labeled. In labeling this data set an attempt was made to label the complete appearance of the object.

There were occasions where the object being labeled were obscured long enough that the location of the object obscured was not clear. In several sequences, specifically some brwncamp and some SBAP sequences vehicles and airplanes are moving away from the camera and recede into the distance. Those objects were labeled as long as the labeler was comfortable saying something was there. It was clear during the testing of ViBe that it was detecting the vehicles after it was extremely difficult to do so by human perception. In evaluating those it may be necessary to consider that some False Negatives are not actually false.

3.0.1 Brown's Town's Camp Sequences

Brown's Town Camp is a campground outside of Bishop, California. Nine DBIR sequences were captured here. The approximate view and camera location can be seen in Fig. 3.1. California Highway 395 is four lanes running North/South at this location. The appoximate location of the camera is marked with a red arrow in Fig. 3.1. The camera Fig. 3.1 is pointing South away from the city of Bishop.

Brwncamp1 captures the scene of a roadway extending in the distance with vehicles vanishing in the distance. Many objects appear in the distance and approach the camera, several also start in the foreground and recede. A roadway enters the scene on the right hand side of the image. Two vehicles enter the main road from the side road. DIfferential vehicle speed with four lanes provided opportunities for objects to occlude fully or partially other objects. The highway veers slightly east, to the image left in Fig. 3.1 approximately 2 km from the position of the camera.

Figure 3.1: Google Maps visible spectrum image of location of data collection as Brown's Town Campground. Accessed 11/11/2023.

Table 3.6 provides a comprehensive list of start and stop frames for all

labeled objects in the sequence. Determining the classification of some objects in brwncamp1 posed challenges, leading to their class ID being set as "Indeterminate." The classification difficulty primarily arose for small objects due to low appearance information, making it challenging for human data labelers to identify their class. Many of these objects remained within the camera's field of view for over 500 frames.

Thumbnails of all objects in brwncamp1 are presented in Figs. 3.2 and 3.2. These thumbnails were automatically generated by identifying the frame with the largest bounding box and extracting the sub-image from the corresponding location. While classifying most of the 15 objects in this sequence was straightforward, objects 13, 14, and 15 presented greater challenges. Object 13 is represented by (y) in LW and (z) in MW, with the MW thumbnail for object 13 being partially obscured by stuck pixels.

Brwncamp2 is very similar to brwncamp1 in terms of camera location and set up. There is only 1 Indeterminate object in this sequence, Object 13. None of the thumbnail images in Fig. 3.3 are significantly impacted by stuck or frozen pixels.

Fig. 3.3 contains a thumbnail image of each object in brwncamp2. Table 3.7 comprehensively lists the objects, start time, stop time, and class in brwncamp2.

Brwncamp3 shares a similar perspective to brwncamp1 and brwncamp2. There are two indeterminate objects, numbers 11 and 12, that come into view from the north at the end of the sequence and given their distance they remain small objects. Target signatures for brwncamp3 are shown in Fig. 3.4.

Table 3.6: Objects Instances with Frame Numbers in Sequence brwncamp1

Sequence	Object	Start	End	Class
brwncamp1	1	1	58	SUV
brwncamp1	2	1	581	Car
brwncamp1	3	1	999	Car
brwncamp1	4	379	960	Car
brwncamp1	5	317	1360	Car
brwncamp1	6	548	1439	Car
brwncamp1	7	504	1157	Car
brwncamp1	8	770	1298	Car
brwncamp1	9	824	1402	SUV
brwncamp1	10	1098	1800	Car
brwncamp1	11	1004	1800	Pickup
brwncamp1	12	1118	1800	Car
brwncamp1	13	1586	1800	Car
brwncamp1	14	1279	1800	Indeterminate
brwncamp1	15	1397	1800	Indeterminate

Table 3.7: Objects Instances with Frame Numbers in Sequence brwncamp2

Sequence	Object	Start	End	Class
brwncamp2	1	1	99	SUV
brwncamp2	2	1	228	Pickup
brwncamp2	3	1	444	SUV and Trailer
brwncamp2	4	60	368	Pickup and Trailer
brwncamp2	5	1	469	Pickup
brwncamp2	6	1	518	Car
brwncamp2	7	1	665	Pickup and Trailer
brwncamp2	8	274	710	Pickup
brwncamp2	9	506	1068	Car
brwncamp2	10	820	1200	Car
brwncamp2	11	277	844	SUV
brwncamp2	12	5	1066	Pickup
brwncamp2	13	981	1200	Indeterminate
brwncamp2	14	1	9	Car

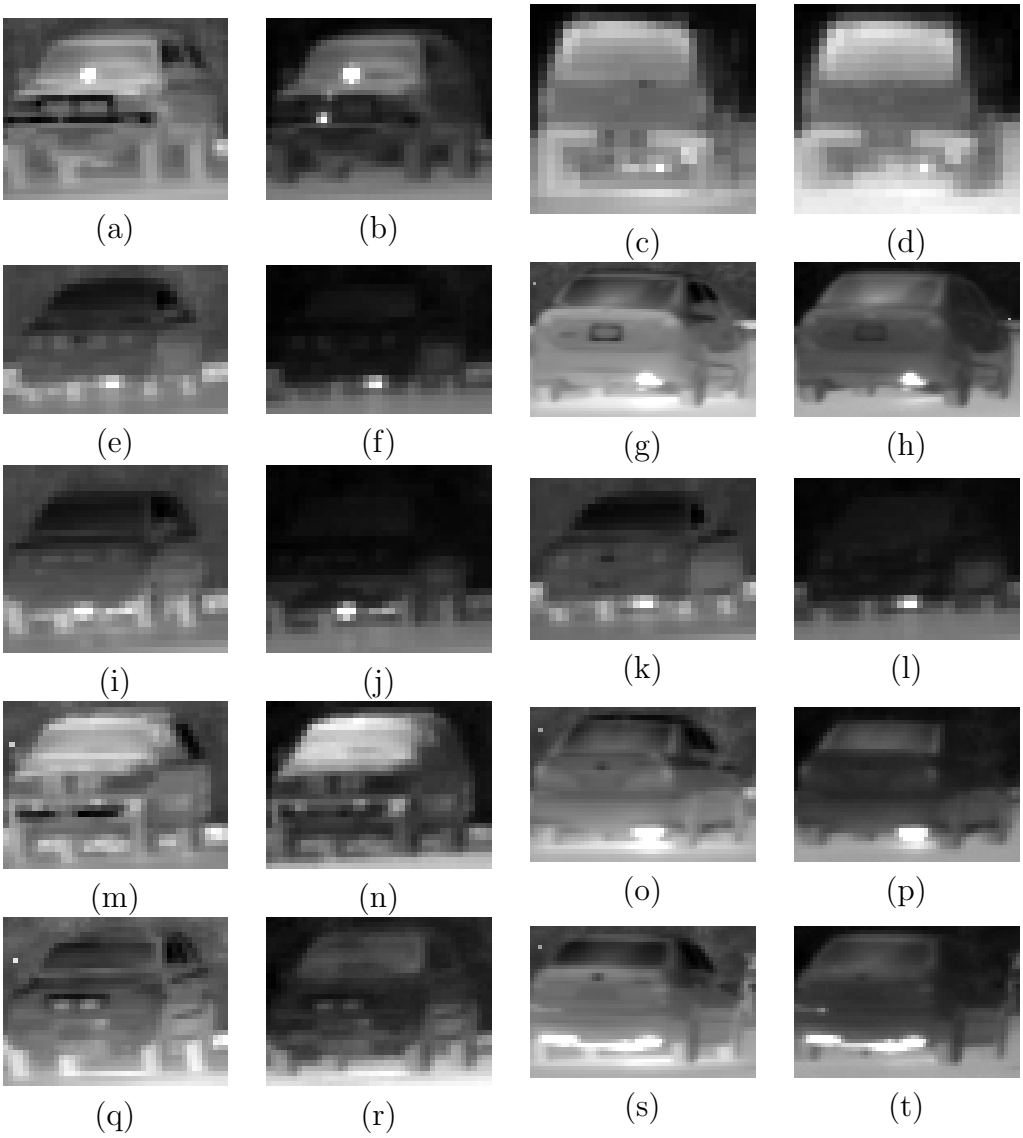


Figure 3.2: Sequence brwncamp1: Target signature. (a) LW Object 1 (b) MW Object 1 (c) LW Object 2 (d) MW Object 2 (e) LW Object 3 (f) MW Object 3 (g) LW Object 4 (h) MW Object 4 (i) LW Object 5 (j) MW Object 5 (k) LW Object 6 (l) MW Object 6 (m) LW Object 7 (n) MW Object 7 (o) LW Object 8 (p) MW Object 8 (q) LW Object 9 (r) MW Object 9 (s) LW Object 10 (t) MW Object 10

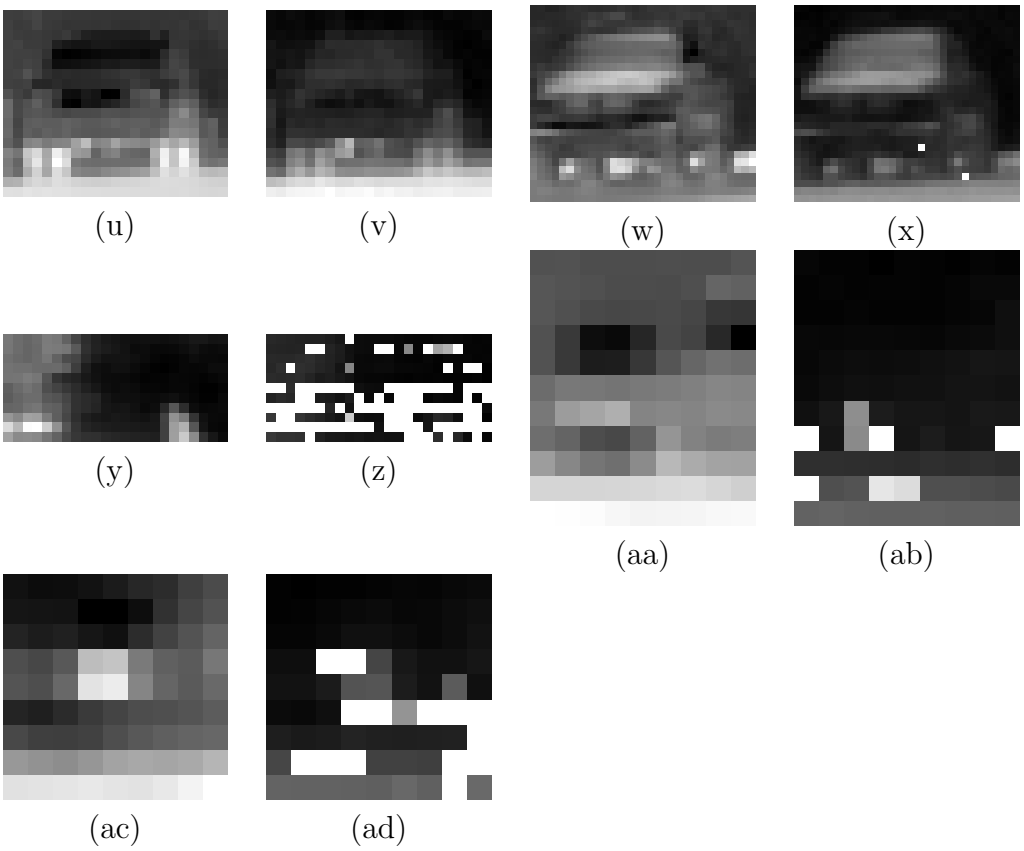


Figure 3.2: Target signatures labeled in sequence brwncamp1. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15

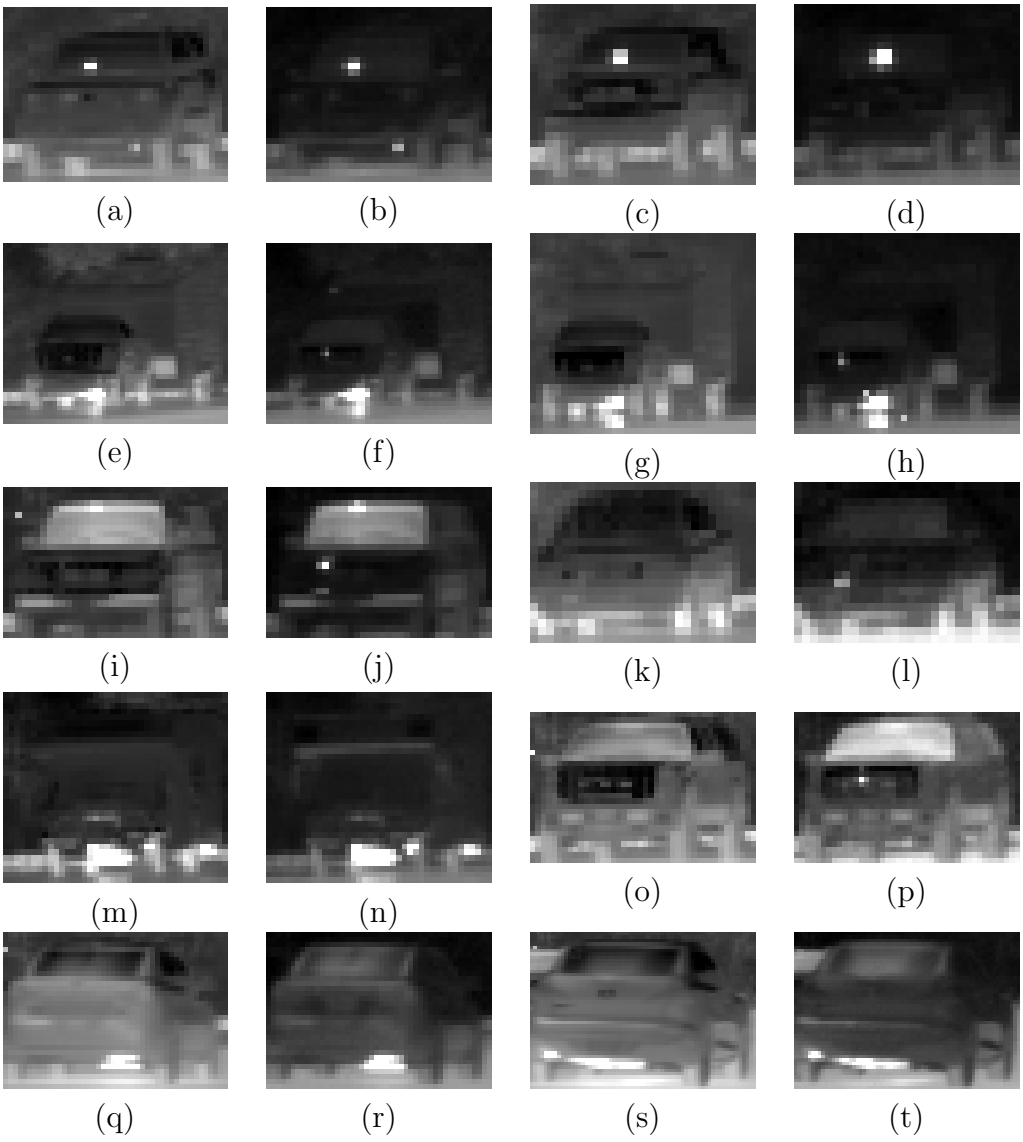


Figure 3.3: Target signatures labeled in sequence brwncamp2. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

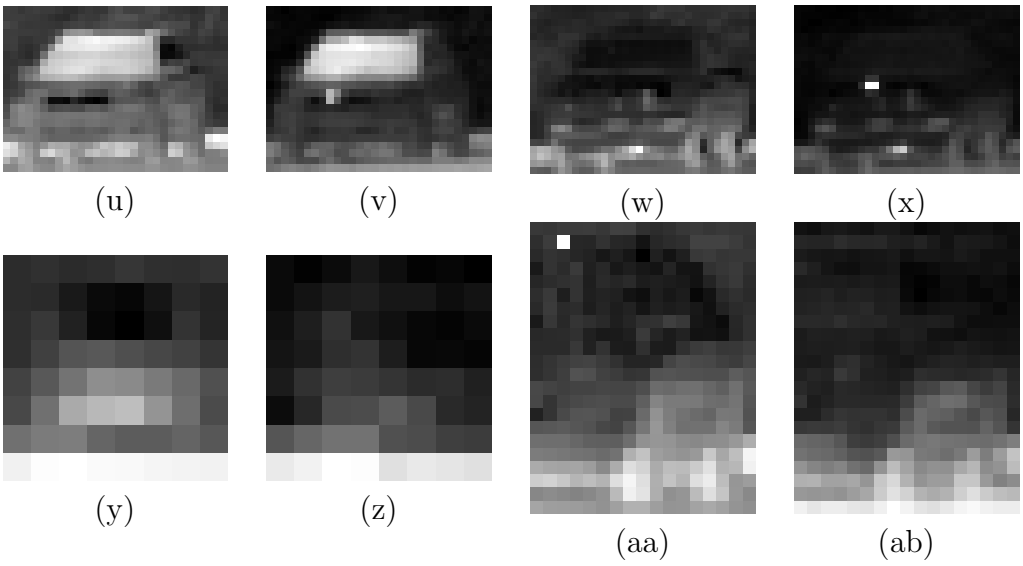


Figure 3.3: Target signatures labeled in sequence brwncamp2. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14

Table 3.8: Objects Instances with Frame Numbers in Sequence brwncamp3

Sequence	Object	Start	End	Class
brwncamp3	1	1	107	Car
brwncamp3	2	148	732	Pickup
brwncamp3	3	386	1031	Van
brwncamp3	4	445	851	Car
brwncamp3	5	418	445	Pickup and Trailer
brwncamp3	6	467	850	Pickup
brwncamp3	7	502	1129	SUV and Trailer
brwncamp3	8	622	1098	SUV
brwncamp3	9	831	1200	Car
brwncamp3	10	809	1200	Car
brwncamp3	11	856	1200	Indeterminate
brwncamp3	12	918	1200	Indeterminate

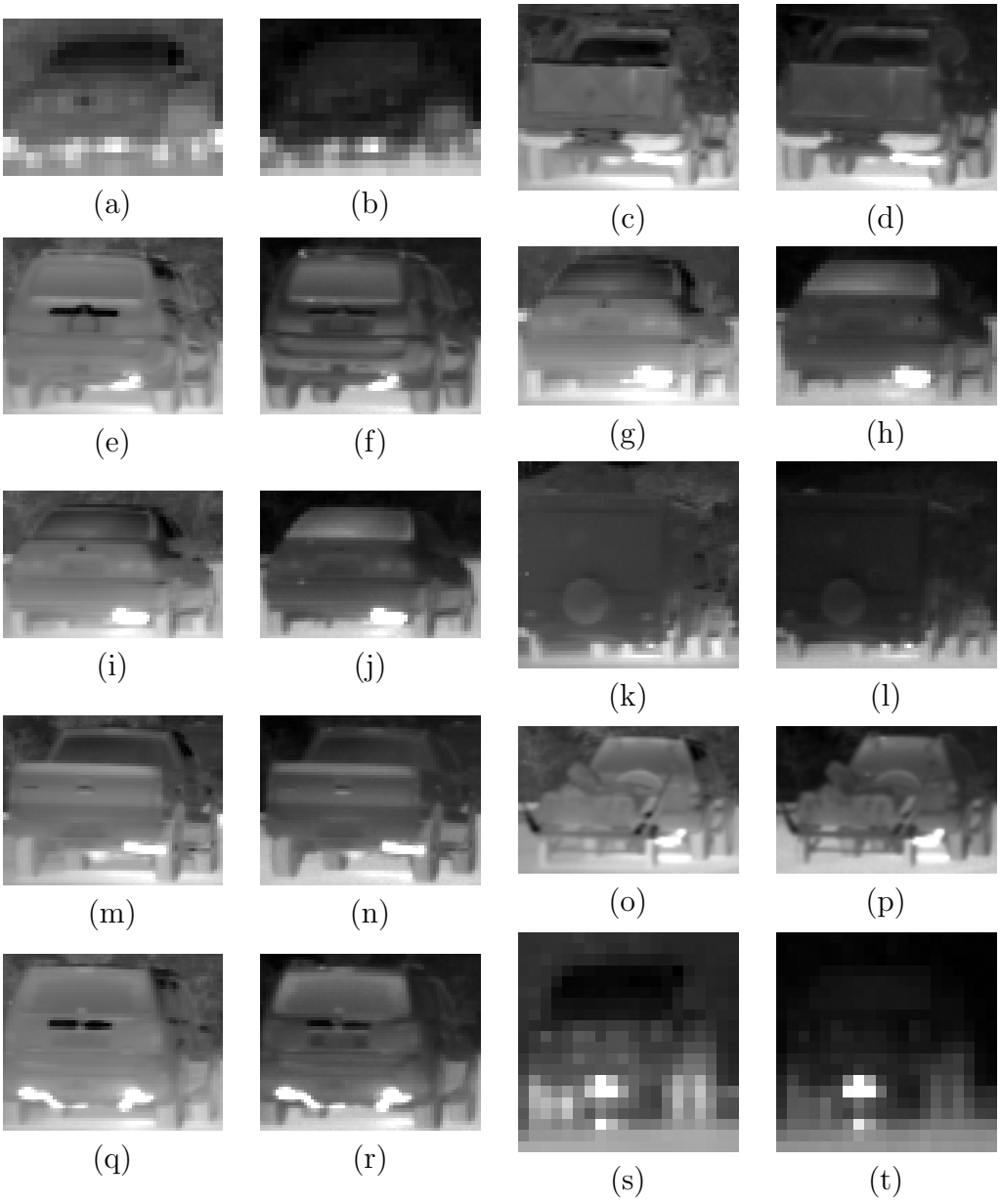


Figure 3.4: Target signatures labeled in sequence brwncamp3. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

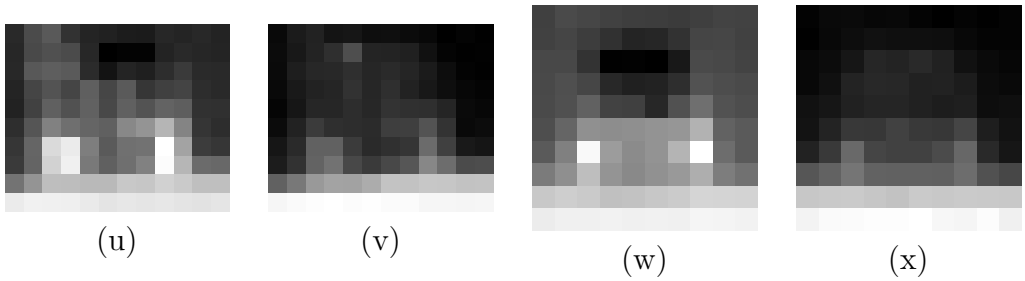


Figure 3.4: Target signatures labeled in sequence brwncamp3. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12

Brwncamp4 marks a substantial shift in perspective from Brwncamps 1-3. The camera initiates at an oblique angle to the roadway, South of East, and pans towards the North. It eventually stops panning, looking approximately North. Notably, the duration that a single object remains within the frame is considerably shorter compared to Brwncamps 1-3, as indicated in Table 3.9.

Brwncamp4 marks a substantial shift in perspective from brwncamps 1-3. The camera initiates at an oblique angle to the roadway, South of East, and pans towards the North. It eventually stops panning, looking approximately North. Notably, the duration that a single object remains within the frame is considerably shorter compared to Brwncamps 1-3, as indicated in Table 3.9.

In Figs. 3.5(c) and (d), parts of a semi-truck are featured. The truck's proximity to the camera results in its appearance exceeding the camera field-of-view, and these thumbnails represent the entire image. This is also the case for Figs. 3.5(e) and (f). However, Figs. 3.5(g), (h), (m), (n), (o), and (p) are partial due to the vehicles being in the nearest south-bound lane, with the lower portion of the vehicles below the lower bound of the camera image.

Figs. 3.5(r) and (t) exhibit significant impact from stuck MW pixels. The sequence involves significant camera motion, presenting challenges for motion-based detection algorithms. Toward the end of the sequence, two indeterminate objects become visible.

Brwncamp5 features a fixed-position camera, pointing slightly East of North. Although the distance at which vehicles come into view is still substantial, it is slightly less than the South view of brwncamp 1-3. An east-west road North of the camera allows vehicles to enter and exit the roadway, creating

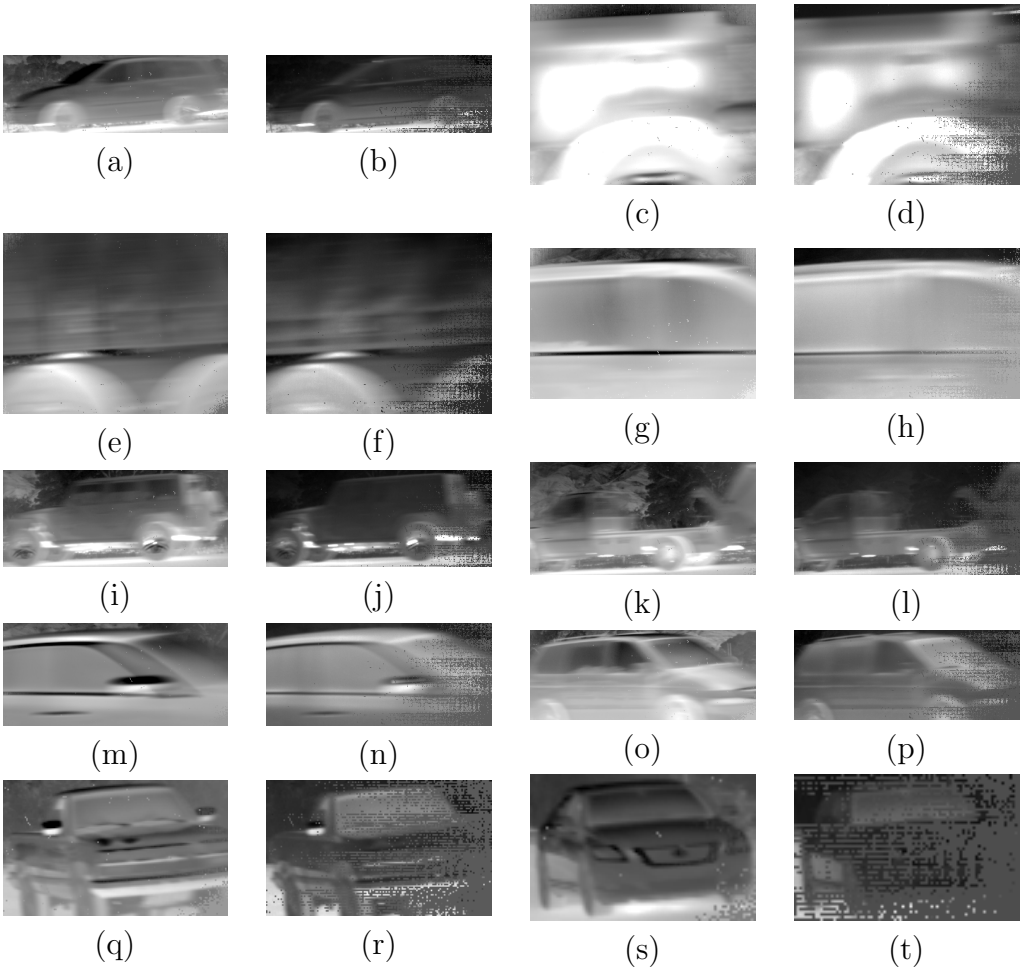


Figure 3.5: Target signatures in sequence brwncamp4. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

Table 3.9: Objects Instances with Frame Numbers in Sequence brwncamp4

Sequence	Object	Start	End	Class
brwncamp4	1	126	137	Car
brwncamp4	2	415	423	Semi and Trailer
brwncamp4	3	422	524	Semi and Trailer
brwncamp4	5	525	534	Car
brwncamp4	6	607	622	SUV
brwncamp4	7	665	682	Pickup
brwncamp4	8	684	695	Car
brwncamp4	9	834	847	Van
brwncamp4	10	1076	1114	Pickup
brwncamp4	11	1116	1200	Car
brwncamp4	12	1123	1200	Indeterminate
brwncamp4	13	1127	1200	Indeterminate

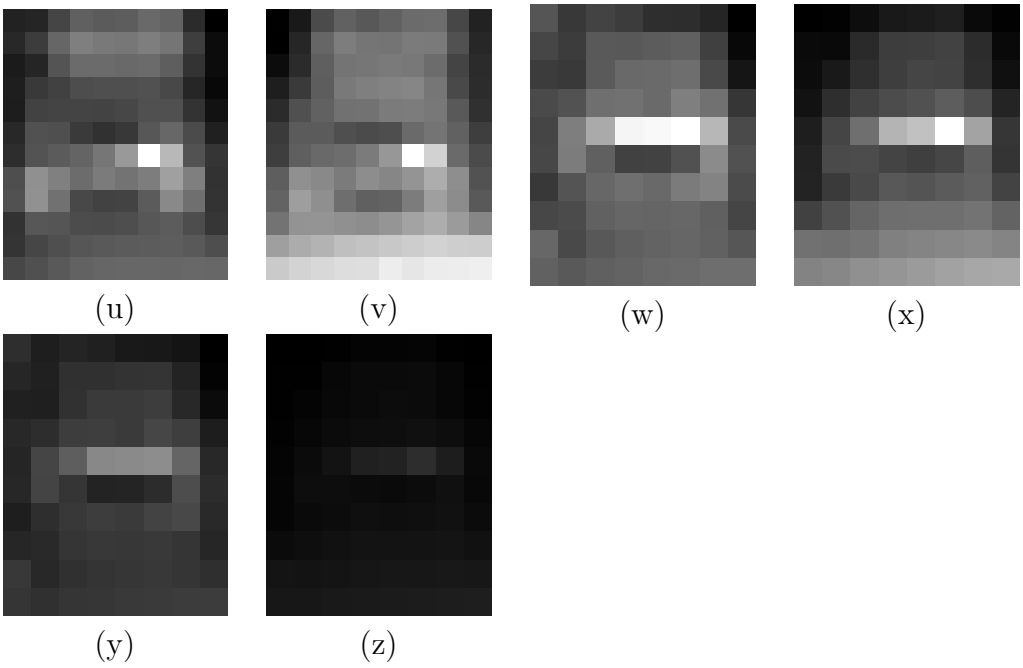


Figure 3.5: Target signatures in sequence brwncamp4. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13

Table 3.10: Objects Instances with Frame Numbers in Sequence brwncamp5

Sequence	Object	Start	End	Class
brwncamp5	1	1	436	Pickup
brwncamp5	2	1	217	Car
brwncamp5	3	1	405	Car
brwncamp5	4	1	250	SUV
brwncamp5	5	119	239	Indeterminate
brwncamp5	6	260	854	Pickup
brwncamp5	7	452	1149	Box Truck
brwncamp5	8	114	735	SUV and Trailer
brwncamp5	9	750	1200	SUV and Trailer
brwncamp5	10	1	906	Semi
brwncamp5	11	813	942	Pickup
brwncamp5	12	655	1200	Pickup
brwncamp5	13	424	708	Car
brwncamp5	14	205	638	Car
brwncamp5	15	1	615	Pickup
brwncamp5	16	1113	1173	Indeterminate
brwncamp5	17	1131	1200	Van
brwncamp5	18	1116	1200	Pickup

several occasions for object occlusion and generation. Tabls 3.10 provides a list of objects, along with the frame in which each object starts getting labeled and the last frame number at which the object is labeled.

Brwncamp5 Object 5, showcased in thumbnails Fig. 3.6(i) and (j), belongs to the indeterminate class. On the other hand, Object 16, presented in thumbnails Fig. 3.6(ae) and (af), is either a Pick-up or SUV so labeled indeterminate.

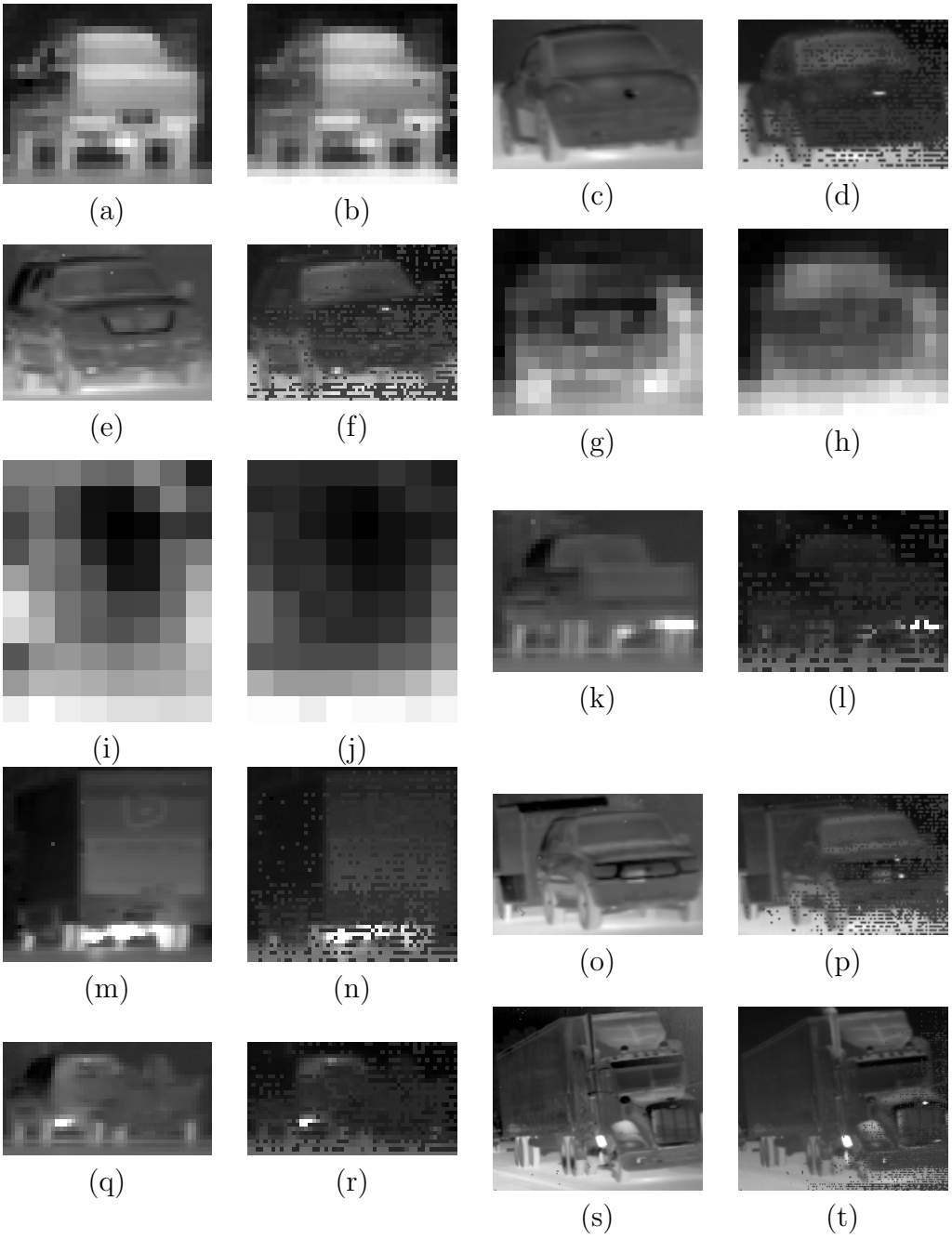


Figure 3.6: Target signatures in sequence brwncamp5. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

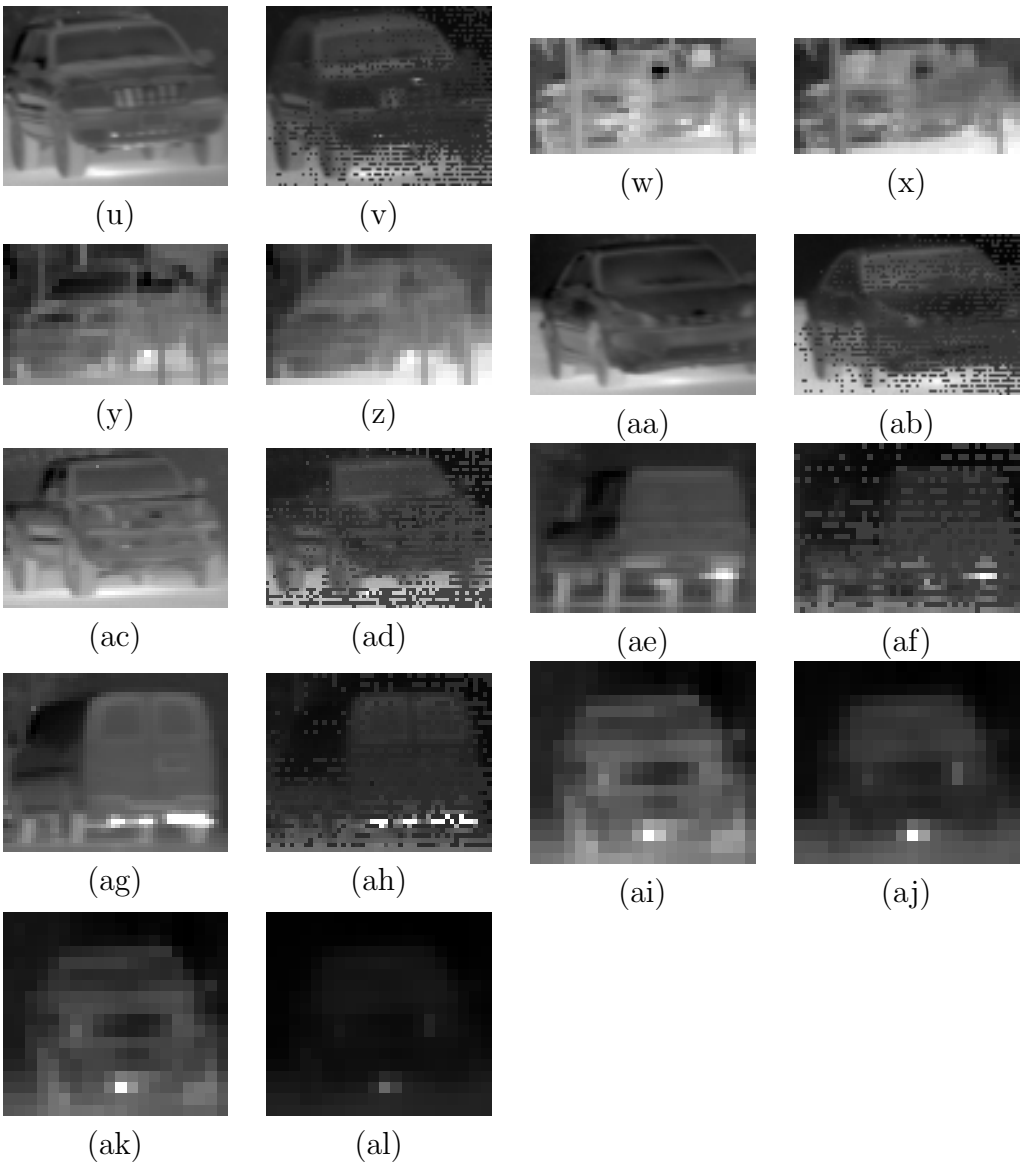


Figure 3.6: Target signatures in sequence brwncamp5. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15; (ae) LW Object 16; (af) MW Object 16; (ag) LW Object 17; (ah) MW Object 17; (ai) LW Object 18; (aj) MW Object 18; (ak) LW Object 19; (al) MW Object 19

Brwncamp6 shares a similar view to Brwncamp5. At the beginning of the sequence, there is a slight camera motion, after which the camera remains stable for the rest of the sequence. This particular sequence is suitable for evaluating recovery from temporary ego-motion of the camera in motion-based detection algorithms.

Table 3.11 provides a comprehensive list of labeled object, start and stop frames, and the object classification. Fig. 3.7 shows a sample target signature for each of these objects.

Objects 5 and 6 are labeled as indeterminate since the sequence starts with them already distant from the camera and moving away. Table 3.11 provides a comprehensive list of objects, including their start and stop frames. Thumbnail images of the objects are presented in Fig. 3.7, with several of the thumbnails showing stuck MW pixels.

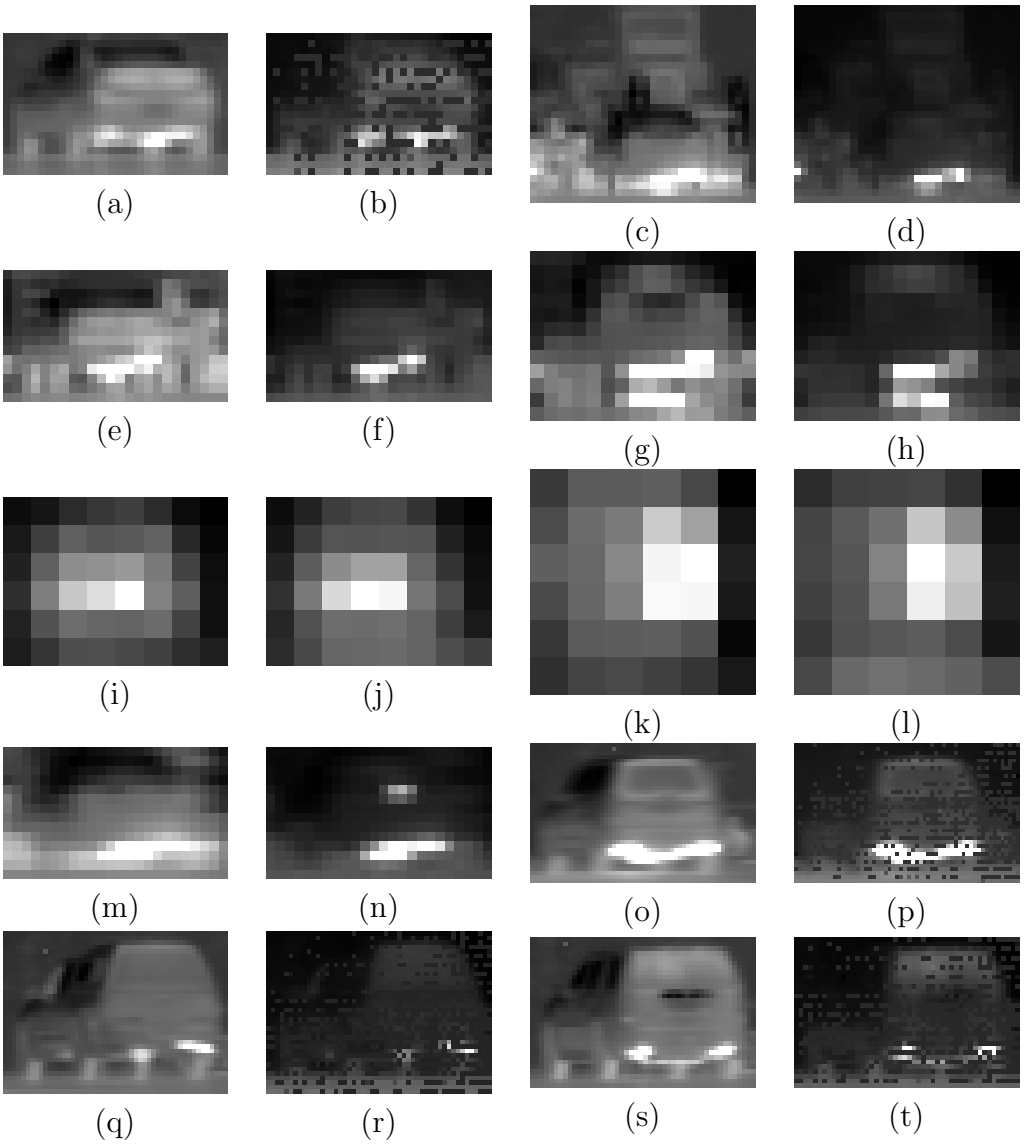


Figure 3.7: Target signatures in sequence brwncamp6. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

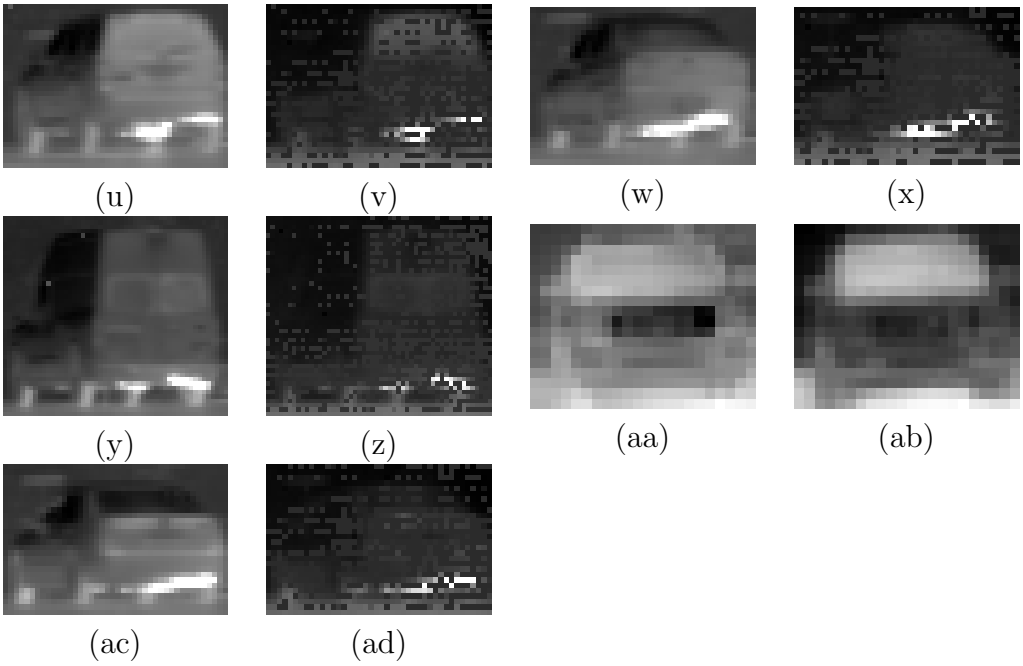


Figure 3.7: Target signatures in Sequence brwncamp6. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15

Table 3.11: Objects Instances with Frame Numbers in Sequence brwncamp6

Sequence	Object	Start	End	Class
brwncamp6	1	1	623	Car
brwncamp6	2	1	623	Semi and Trailer
brwncamp6	3	1	375	Car
brwncamp6	4	1	205	Car
brwncamp6	5	1	238	Indeterminate
brwncamp6	6	1	9	Indeterminate
brwncamp6	7	1	373	Car
brwncamp6	8	6	511	SUV
brwncamp6	9	17	801	Pickup
brwncamp6	10	384	1053	SUV
brwncamp6	11	609	1200	Pickup
brwncamp6	12	730	1200	SUV
brwncamp6	13	1020	1200	SUV
brwncamp6	14	1069	1200	Car
brwncamp6	15	1125	1200	Van

Brwncamp7 and brwncamp6 share a similar camera setup and view. While brwncamp6 has initial ego-motion, brwncamp7 does not, and the camera remains stationary throughout the sequence.

Table 3.12 lists the objects, start frame, stop frame, and classification for the objects in brwncamp7. Fig. 3.8 lists the thumbnails for brwncamp7 target signatures.

Table 3.12: Objects Instances with Frame Numbers in Sequence brwncamp7

Sequence	Object	Start	End	Class
brwncamp7	1	1	55	SUV
brwncamp7	2	1	77	SUV
brwncamp7	3	1	512	Pickup
brwncamp7	4	1	95	Car
brwncamp7	5	1	180	Car
brwncamp7	6	1	219	Car
brwncamp7	7	1	285	Pickup
brwncamp7	8	1	367	Indeterminate
brwncamp7	9	99	416	Box Truck
brwncamp7	10	333	781	Pickup
brwncamp7	11	502	732	Car
brwncamp7	12	683	1000	Pickup
brwncamp7	13	773	1200	Car
brwncamp7	14	938	1200	Pickup
brwncamp7	15	1001	1200	Pickup
brwncamp7	16	1032	1200	Pickup
brwncamp7	17	1062	1200	Pickup?

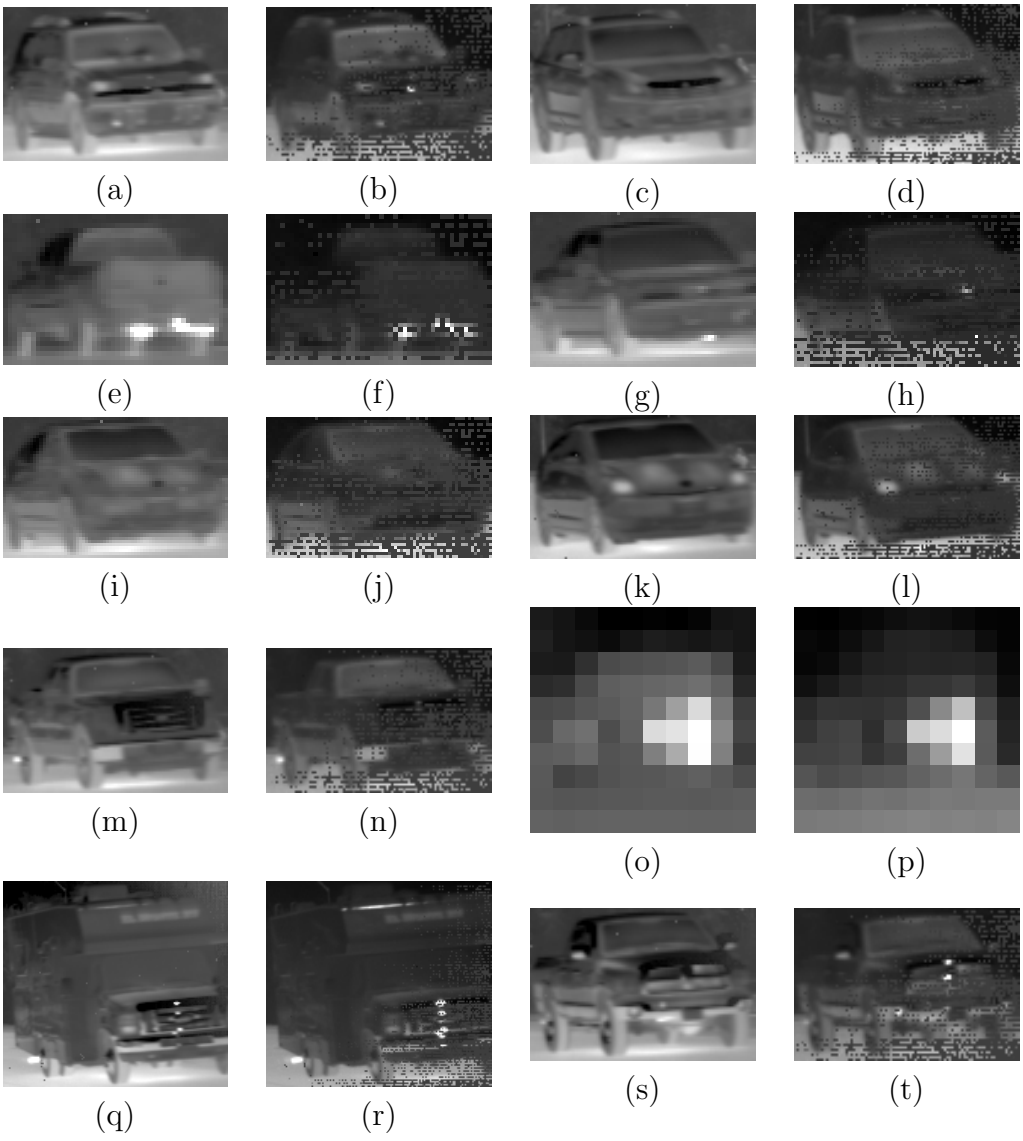


Figure 3.8: Target signatures in sequence brwncamp7. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

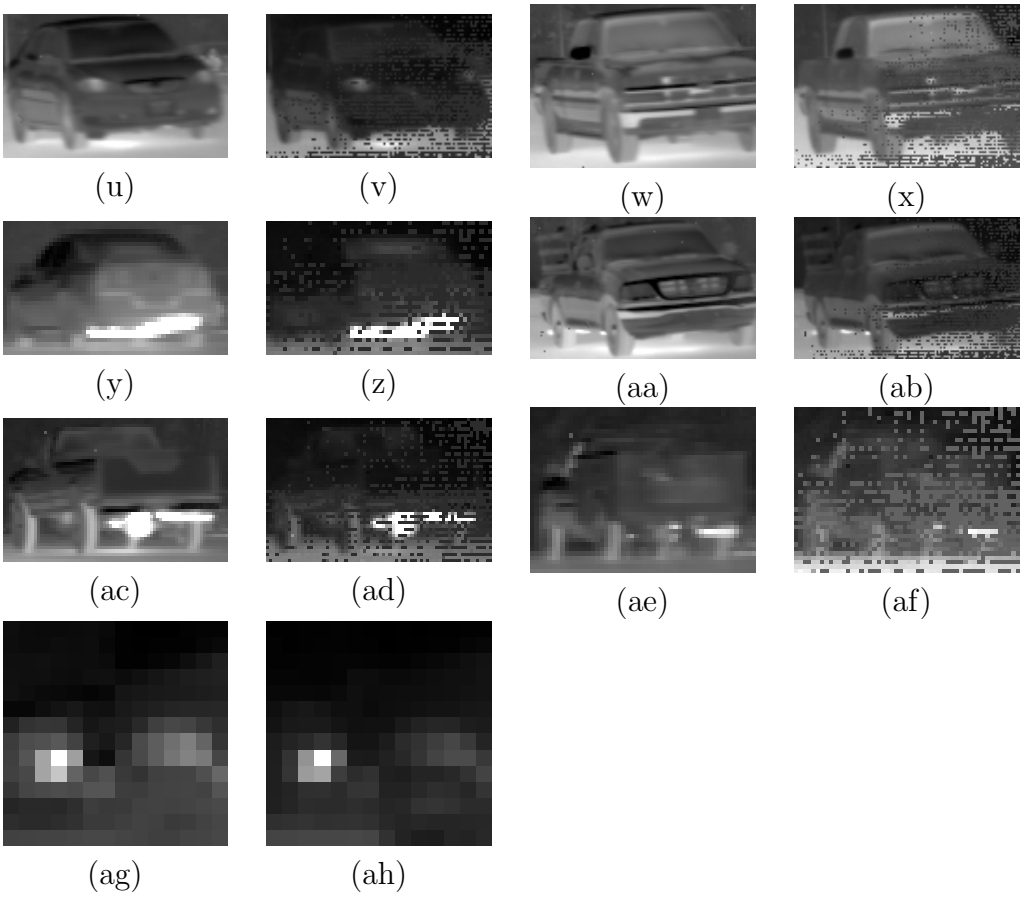


Figure 3.8: Target signatures in sequence brwncamp7. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15; (ae) LW Object 16; (af) MW Object 16; (ag) LW Object 17; (ah) MW Object 17

Brwncamp8 shares a similar view with brwncamp6 and brwncamp7. As the sequence begins, there are two indeterminate objects (objects 4 and 5 in Table 3.13) receding into the distance. Additionally, brwncamp8 features two people, objects 1 and 2, who are approaching the camera throughout the entire 1,200-frame sequence. Thumbnails of these individuals can be seen in Fig. 3.9(a-d), with Object 1 identified as Joseph Havlicek, my adviser, and Object 2 myself. Table 3.13 comprehensively lists the object information for brwncamp8, with Fig. 3.9 providing examples of appearance information.

Brwncamp9 is similar to brwncamp4 with the perspective of the camera being closer to perpendicular to the roadway. This resulted in the shorter duration object presences seen in Table 3.14. The perspective also resulted in several partial images of objects as seen in Figs. 3.9 (e)-(ah). There is also significant motion blur from both the motion of the camera and the vehicles.

Table 3.13: Objects Instances with Frame Numbers in Sequence brwncamp8

Sequence	Object	Start	End	Class
brwncamp8	1	1	1200	Person
brwncamp8	2	1	1200	Person
brwncamp8	3	1	399	Indeterminate
brwncamp8	4	1	356	Pickup
brwncamp8	5	1	330	Indeterminate
brwncamp8	6	6	463	SUV
brwncamp8	7	443	647	SUV
brwncamp8	8	329	669	Pickup
brwncamp8	9	454	826	Car
brwncamp8	10	390	844	SUV
brwncamp8	11	648	1113	Pickup
brwncamp8	12	667	1054	Pickup
brwncamp8	13	699	1127	Pickup
brwncamp8	14	755	1155	Car
brwncamp8	15	821	1169	Pickup
brwncamp8	16	852	1186	Pickup
brwncamp8	17	663	938	Car
brwncamp8	18	559	1066	Car
brwncamp8	19	595	1118	Pickup
brwncamp8	20	1147	1200	Pickup
brwncamp8	21	1025	1200	Pickup
brwncamp8	22	1129	1200	Car

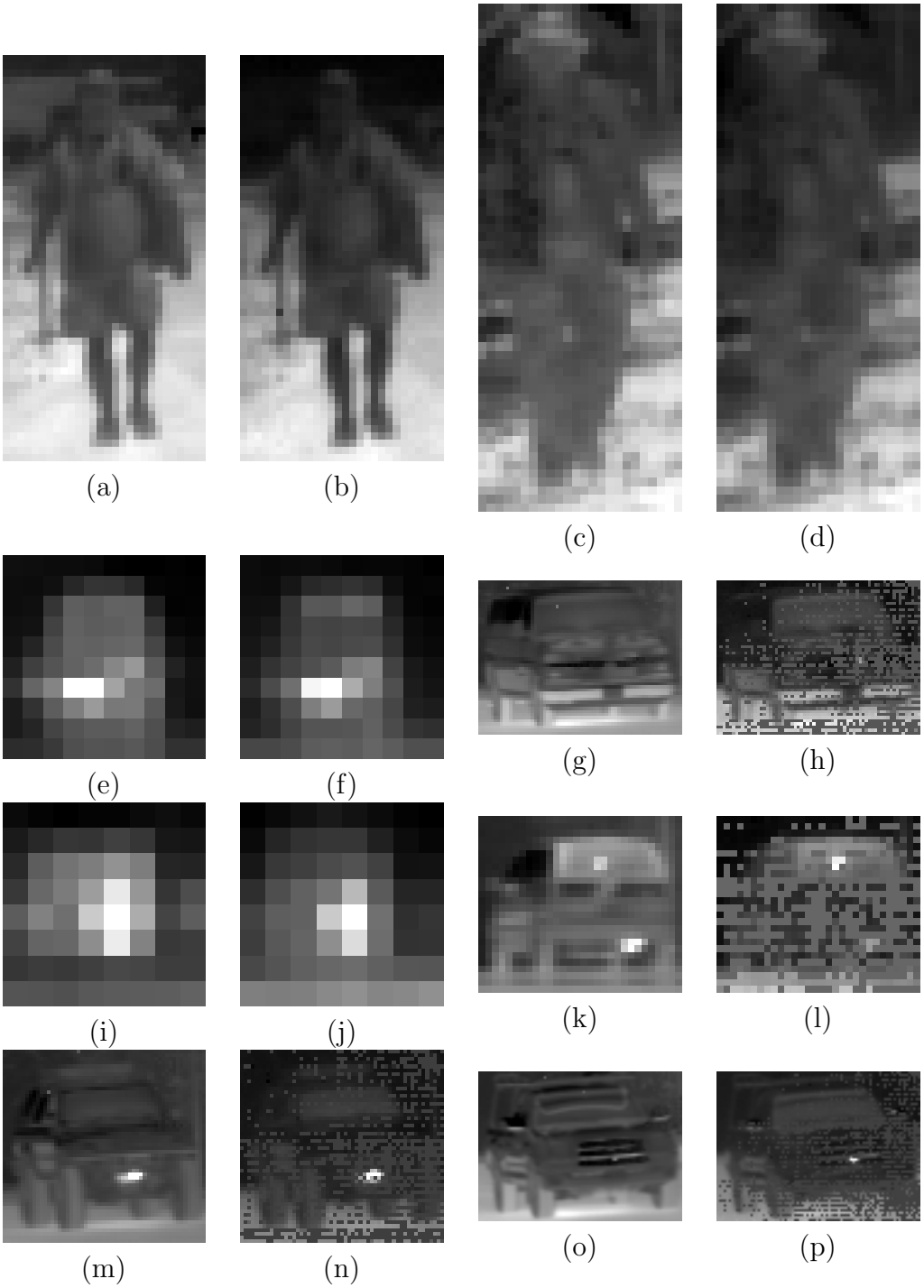


Figure 3.9: Target signatures in sequence brwncamp8. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8

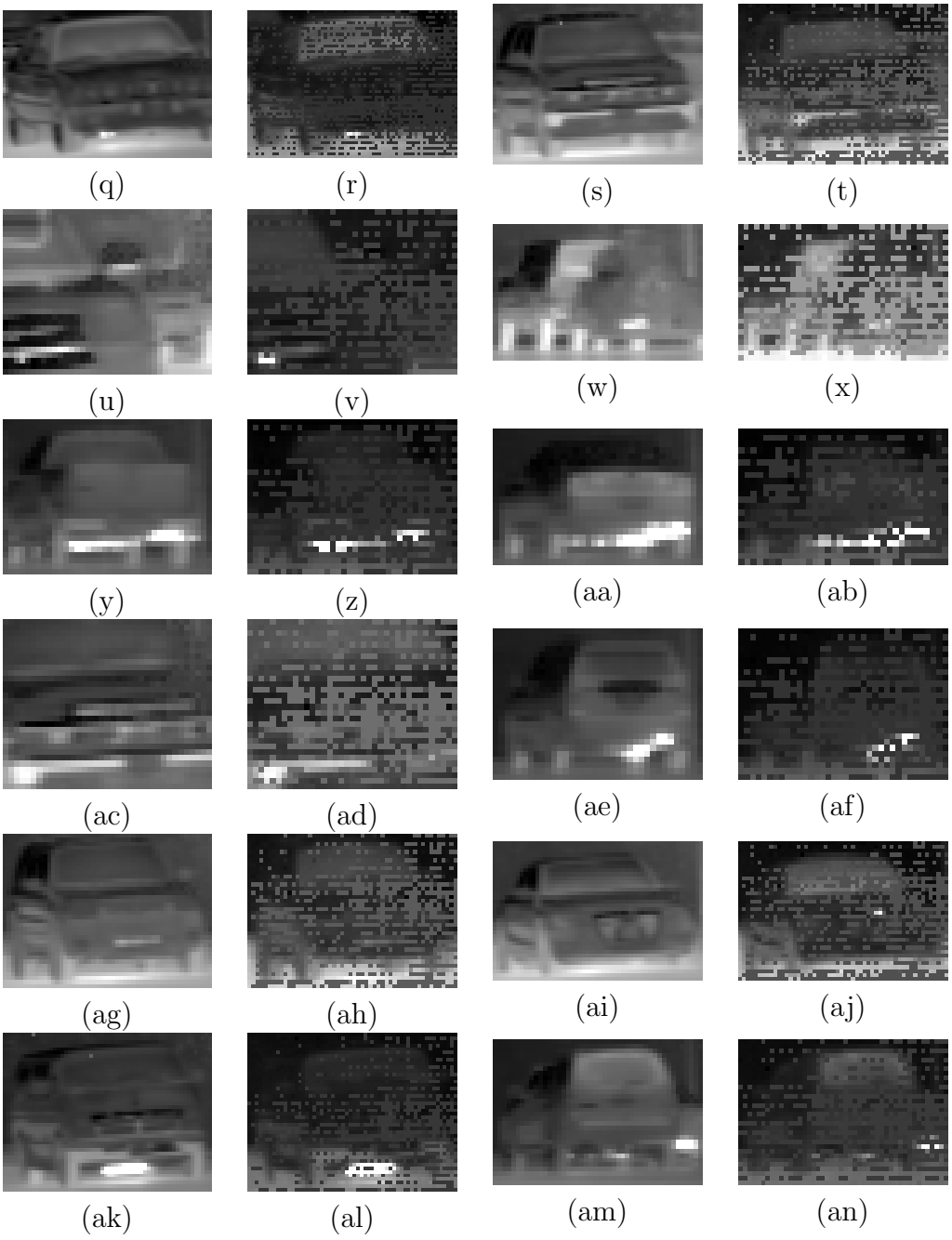
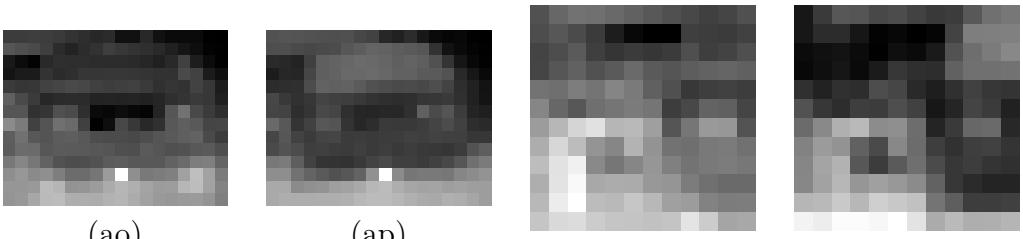


Figure 3.9: Target signatures in sequence brwncamp8. (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10; (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15; (ae) LW Object 16; (af) MW Object 16; (ag) LW Object 17; (ah) MW Object 17; (ai) LW Object 18; (aj) MW Object 18; (ak) LW Object 19; (al) MW Object 19; (am) LW Object 20; (an) MW Object 20



(ao) (ap) (aq) (ar)

Figure 3.9: Target signatures in sequence brwncamp8. (ao) LW Object 21; (ap) MW Object 21; (aq) LW Object 22; (ar) MW Object 22

Table 3.14: Objects Instances with Frame Numbers in Sequence brwncamp9

Sequence	Object	Start	End	Class
brwncamp9	1	26	44	Pickup
brwncamp9	2	174	193	Pickup
brwncamp9	3	317	421	SUV
brwncamp9	5	426	461	Semi and Trailer
brwncamp9	6	469	478	SUV
brwncamp9	7	508	514	Car
brwncamp9	8	573	578	Car
brwncamp9	9	701	711	Car
brwncamp9	10	846	859	Car
brwncamp9	11	859	865	SUV
brwncamp9	12	918	926	Pickup
brwncamp9	13	978	989	Pickup
brwncamp9	14	994	1011	Car
brwncamp9	15	1075	1083	Car
brwncamp9	16	1099	1105	Car
brwncamp9	17	1182	1194	Car

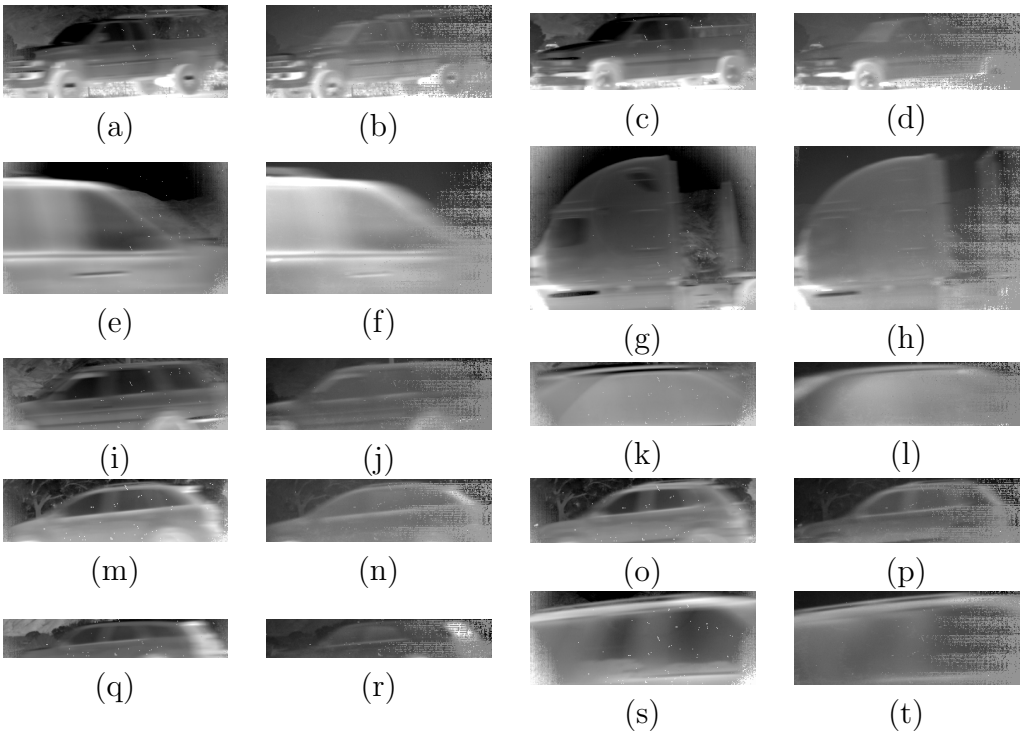


Figure 3.10: Target signatures in sequence brwncamp9. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10

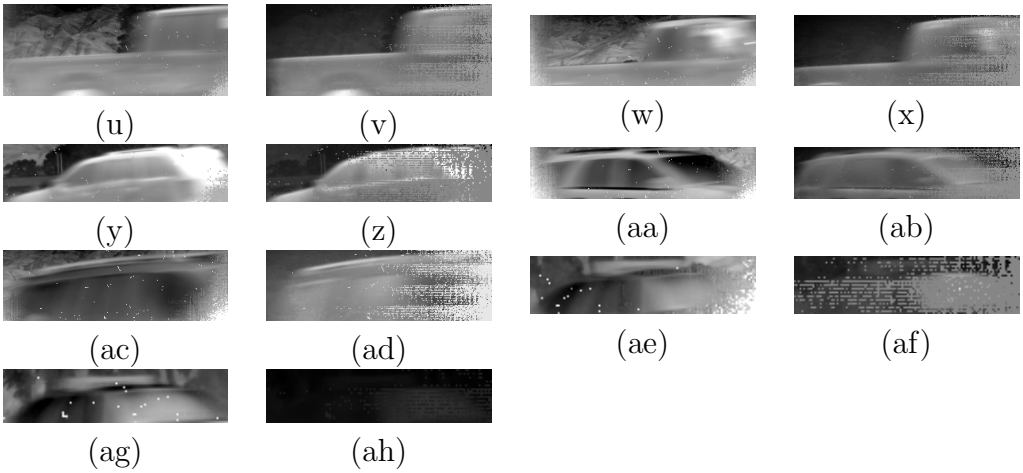


Figure 3.10: Target signature in sequence brwncamp9. (u) LW Object 11; (v) MW Object 11; (w) LW Object 12; (x) MW Object 12; (y) LW Object 13; (z) MW Object 13; (aa) LW Object 14; (ab) MW Object 14; (ac) LW Object 15; (ad) MW Object 15; (ae) LW Object 16; (af) MW Object 16; (ag) LW Object 17; (ah) MW Object 17

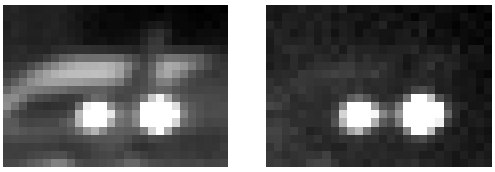
3.0.2 Santa Barbara Airport Sequences

Santa Barbara Airport (SBAP) sequences were collected at Santa Barbara Airport, in Goletta, CA. The lower part of the view often contains a chain-link fence to prevent civilians getting on the runway or the road that encircles the runways. In these sequences instances of airplanes launching, taxiing, and landing were captured. Several sequences included vehicles on the road that passes behind the airport, fuel trucks, and service trucks.

SBAP1 is a 160 frame sequence featuring the launch of a single jet aircraft as in Table 3.15. The two bright spots in Fig. 3.11(a) and (b) are the thermal signatures of the jet engines in MWIR and LWIR. The camera is not stationary during this sequence.

Table 3.15: Objects Instances with Frame Numbers in Sequence SBAP1

Sequence	Object	Start	End	Class
SBAP1	1	1	160	Airplane



(a) (b)

Figure 3.11: Target signatures in Sequence SBAP1. (a) LW Object 1; (b) MW Object 1

The first 59 frames SBAP2 contain an aircraft, with a view of the sky as background. The camera then pans down and captures three birds, a car, and a pickup. For frames see Table 3.16. The birds tended to be small and distant but their general outlines can be infered from the thumbnails in Figs. 3.12(c)-(f) and (i)-(j). It is challenging to discern the bird from the cluttered background in the MWIR in Figs. 3.12 (d), (f), and (j). There is significant ego-motion of the camera throughout this sequence. Calibration issues are apperent in LWIR in Fig. 3.12 (g) and stuck pixels impact the target signatures displayed in Figs. 3.12 (h), (l),and (n). The MWIR signature of the aircraft in Fig. 3.12 (b) is faint and hard to distinguish the Object 1 from background in the thumbnail.

Table 3.16: Objects Instances with Frame Numbers in Sequence SBAP2

Sequence	Object	Start	End	Class
SBAP2	1	1	59	Airplane
SBAP2	2	627	630	Bird
SBAP2	3	638	679	Bird
SBAP2	5	779	855	Car
SBAP2	6	808	822	Bird
SBAP2	7	882	953	Pickup

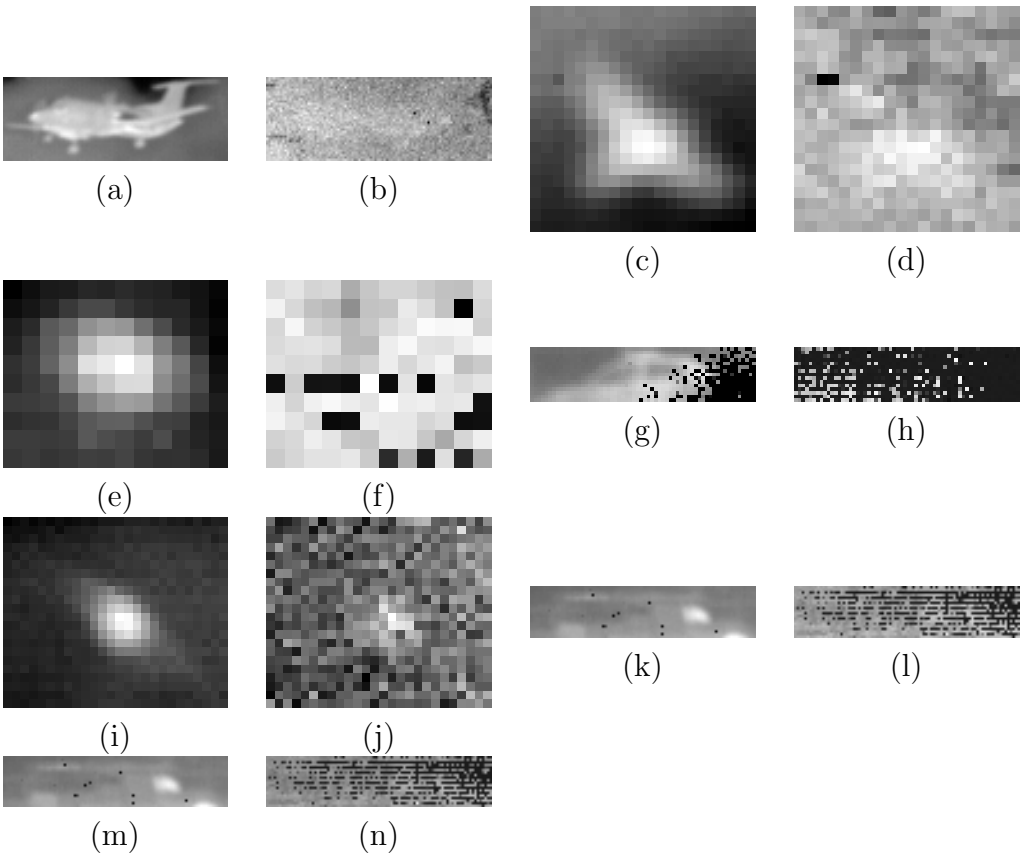


Figure 3.12: Target signatures in sequence SBAP2. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7

Table 3.17: Objects Instances with Frame Numbers in Sequence SBAP3-9

Sequence	Object	Start	End	Class
SBAP3	1	1	479	Fuel Truck
SBAP3	2	104	140	Bird
SBAP5	1	1	640	Airplane
SBAP5	2	513	571	Fuel Truck
SBAP6	1	1	639	Airplane
SBAP7	1	1	639	Airplane
SBAP8	1	1	639	Airplane
SBAP8	2	38	117	Bird
SBAP9	1	1	639	Airplane
SBAP9	2	541	639	Airplane

In SBAP3, two objects are being labeled. The first object is a fuel truck transiting the perimeter road, as depicted in Fig. 3.13 (a) and (b). The camera follows this truck as it approaches and then turns in front of it, inducing significant ego-motion of the camera. Object 2 is a bird, seen in Fig. 3.13 (c) and (d). It's worth noting that all birds in this dataset exhibit low appearance information. While the outline is recognizable to a human in the LWIR, distinguishing the bird from the background in MWIR proves to be challenging. Frame information is provided in Table 3.17, and thumbnails of the objects' appearance in MWIR and LWIR are presented in Fig. 3.13. SBAP4 has no targets and is not discussed in this section. SBAP5 features an aircraft taxiing, labeled as Object 1. There is a significant change in appearance as the aircraft turns to align with the runway before take-off. Briefly, a fuel truck, labeled as Object 2, occludes the aircraft. The frames for the objects in this sequence are detailed in Table 3.17. Sample object appearances are in Fig. 3.13 (c) - (f).

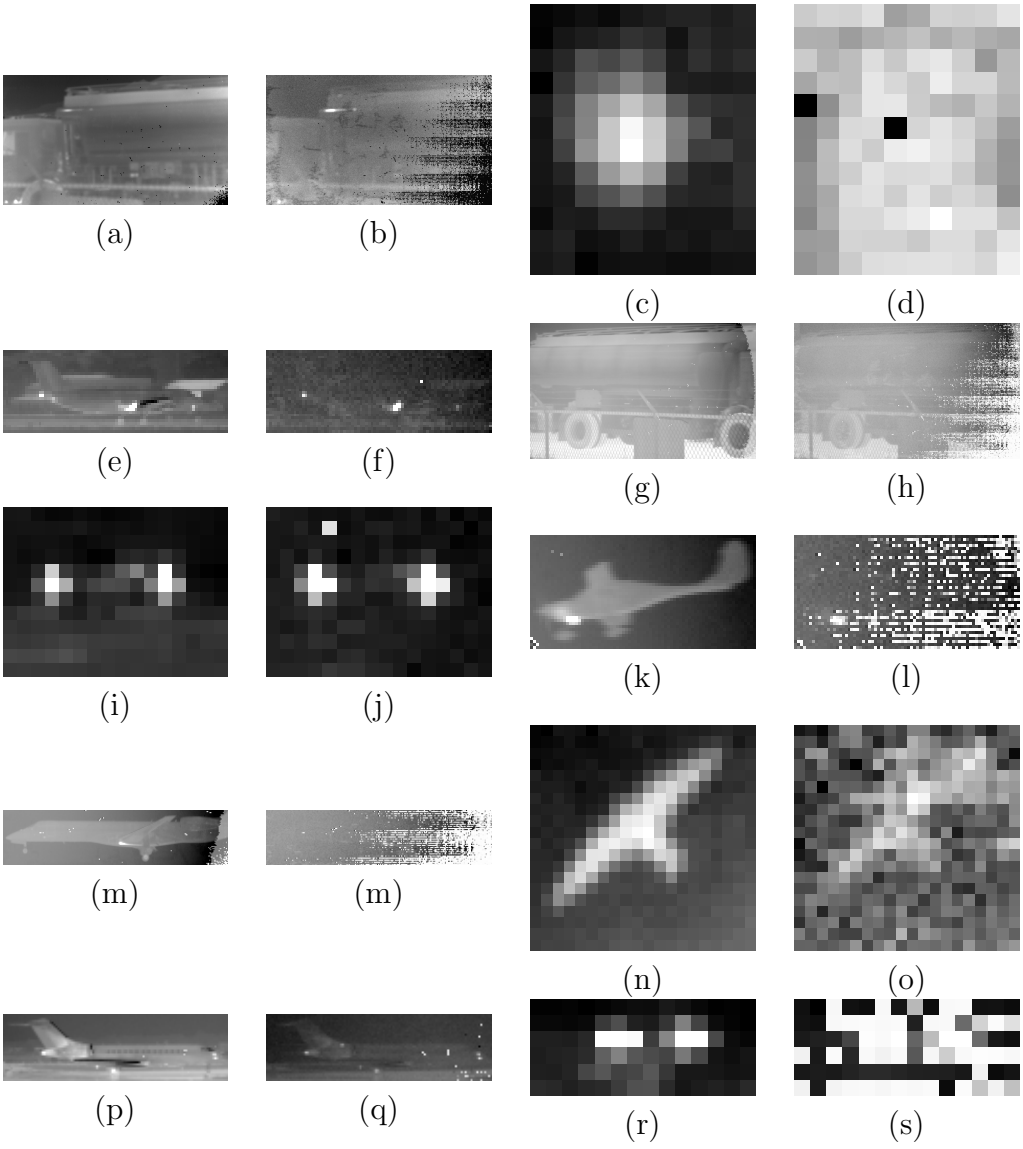


Figure 3.13: Target signatures in Sequence SBAP3-6. (a) SBAP3 LW Object 1; (b) SBAP3 MW Object 1; (c) SBAP3 LW Object 2; (d) SBAP3 MW Object 2; (e) SBAP5 LW Object 1; (f) SBAP5 MW Object 1; (g) SBAP5 LW Object 2; (h) SBAP5 MW Object 2; (i) SBAP6 LW Object 1; (j) SBAP6 MW Object 1; (j) SBAP7 LW Object 1; (k) SBAP7 MW Object 1; (l) SBAP8 LW Object 1; (m) SBAP8 MW Object 1; (n) SBAP8 LW Object 2; (o) SBAP8 MW Object 2; (p) SBAP9 LW Object 1; (q) SBAP9 MW Object 1; (r) SBAP9 LW Object 2; (s) SBAP9 MW Object 2

SBAP6 is a sequence of the final portion of the take-off of an aircraft. The aircraft ascends and continues to become smaller and more distant to the camera. In the final frames the view is obscured as a person walks in front of the camera. Frame information is provided in Table 3.17 and the engine signatures can be seen in Figs. 3.13 (i) and (j).

SBAP7 follows the landing of a propeller driven aircraft. The initial portion of the sequence has clouds as the background. As the airplane descends mountains and buildings enter the sequence. The camera pans significantly and there is a significant ego-motion. The frame information for the aircraft can be seen in Table 3.17 and thumbnails in Fig. 3.13 (k) and (l).

SBAP8 shows the landing of a jet aircraft much larger than the propeller aircraft seen in SBAP7. A bird is also labeled in SBAP8. There is significant ego-motion in this sequence. Frame information is provided in Table 3.17 and thumbnails in Fig. 3.13 (l) - (o).

SBAP9 is as sequence of the slow taxi of a large jet engined airplane. A second aircraft transits in the background as part of a take-off sequence. The profile of the first aircraft are shown in the thumbnails Fig. 3.13 (p) and (q). The signature of the two jet engines are presented as thumbnails in Fig. 3.13 (r) and (s). Fig. 3.13 (s) is significantly effected by stuck pixels. Frame information for this sequence is in Table 3.17.

SBAP10 is a sequence with vehicles and birds as labeled objects. There five birds and 4 vehicles. There are intermittent repositionings of the camera throughout the sequence. Frozen pixels can be seen in the LWIR images as seen in Fig. 3.14 (e), (m), (s), and (u). The avian images are relatively small

Table 3.18: Objects Instances with Frame Numbers in Sequence SBAP10

Sequence	Object	Start	End	Class
SBAP10	1	5	26	Bird
SBAP10	2	18	34	Bird
SBAP10	3	16	92	Pickup
SBAP10	4	51	76	Bird
SBAP10	5	58	73	Bird
SBAP10	6	83	128	Bird
SBAP10	7	258	327	Car
SBAP10	8	412	449	Bird
SBAP10	9	441	472	Bird
SBAP10	10	558	634	Pickup
SBAP10	11	594	639	Car

and somewhat hard to make the outline of. Frame information for the vehicles and birds is in Table 3.18.

SBAP11 follows the departure of an aircraft. The jet engine signatures as seen in Fig. 3.15 are visible in MWIR and LWIR for the first 350 frames of the sequence as reported in Table 3.19. The camera was maneuvered to keep the aircraft in the field of view. It may be possible to extend the labeling procedure with enhancement techniques.

SBAP12 is a sequence of images of two vehicles traveling from the right of the frame to the left. The vehicle's MWIR and LWIR signatures can be examined in Fig. 3.15(c)-(f) with start/stop frames in Table 3.19.

SBAP13 has a similar view approximately perpindicular to the road, in this sequence the vehicles are moving left to right. Signatures in Fig. 3.15 (g)-(n). Two birds are also visible in this sequence. Start/stop time for these vehicles and birds are available in Table 3.19.

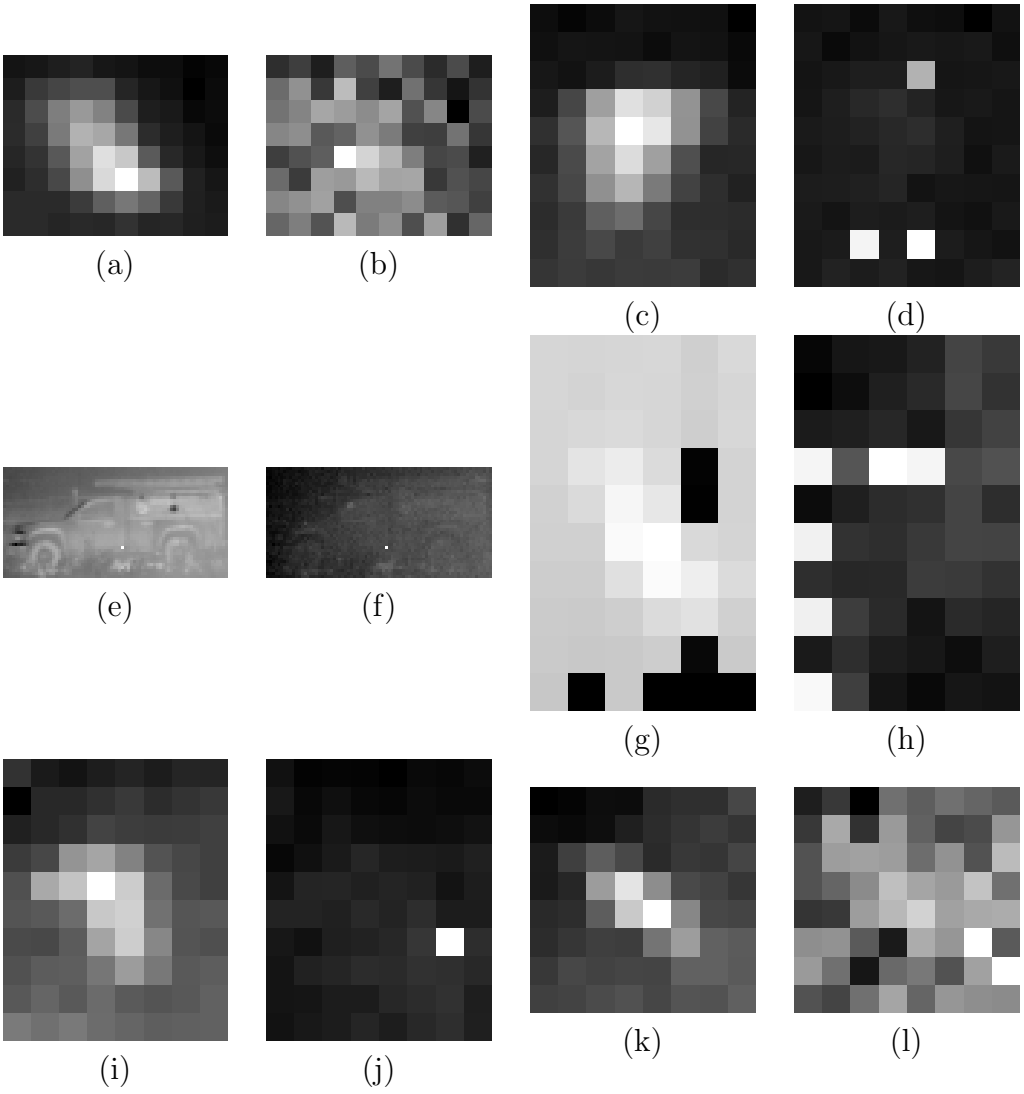


Figure 3.14: Target signatures in sequence SBAP10. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6

SBAP14 is the taxiing of jet aircraft, thumbails available in Fig. 3.15 (o) and (p). Frame start/stop available in Table 3.19. SBAP15 continues the taxiing of the aircraft seen in SBAP14. There is a discontinuity between the sequences. The differential emissivity in IR is visible in the thumbnails and

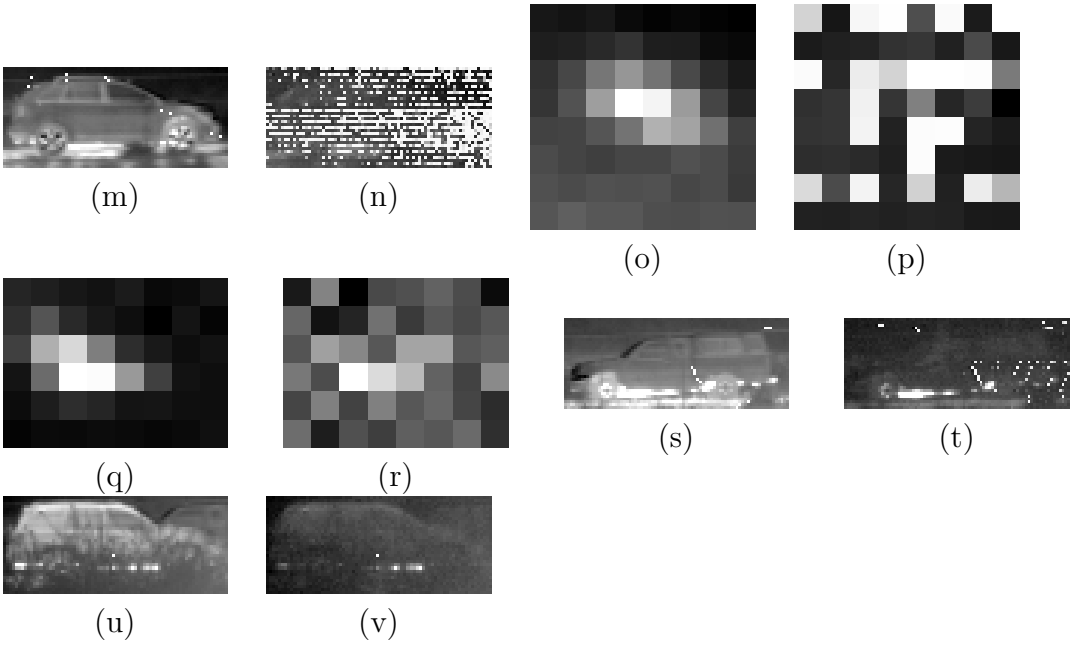


Figure 3.14: Target signatures in sequence SBAP10. (m) LW Object 7; (n) MW Object 7; (o) LW Object 8; (p) MW Object 8; (q) LW Object 9; (r) MW Object 9; (s) LW Object 10; (t) MW Object 10; (u) LW Object 11; (v) MW Object 11

"Frontier" Airlines logo is visible in Fig. 3.15 (q) and (r). Start/stop frames for SBAP15 are in Table 3.19.

Table 3.19: Objects Instances with Frame Numbers in Sequence SBAP11-15

Sequence	Object	Start	End	Class
SBAP11	1	1	350	Airplane
SBAP12	1	1	64	Pickup
SBAP12	2	38	103	Car
SBAP13	1	12	75	Car
SBAP13	2	107	119	Bird
SBAP13	3	167	235	Car
SBAP13	4	171	190	Bird
SBAP14	1	1	319	Airplane
SBAP15	1	1	319	Airplane

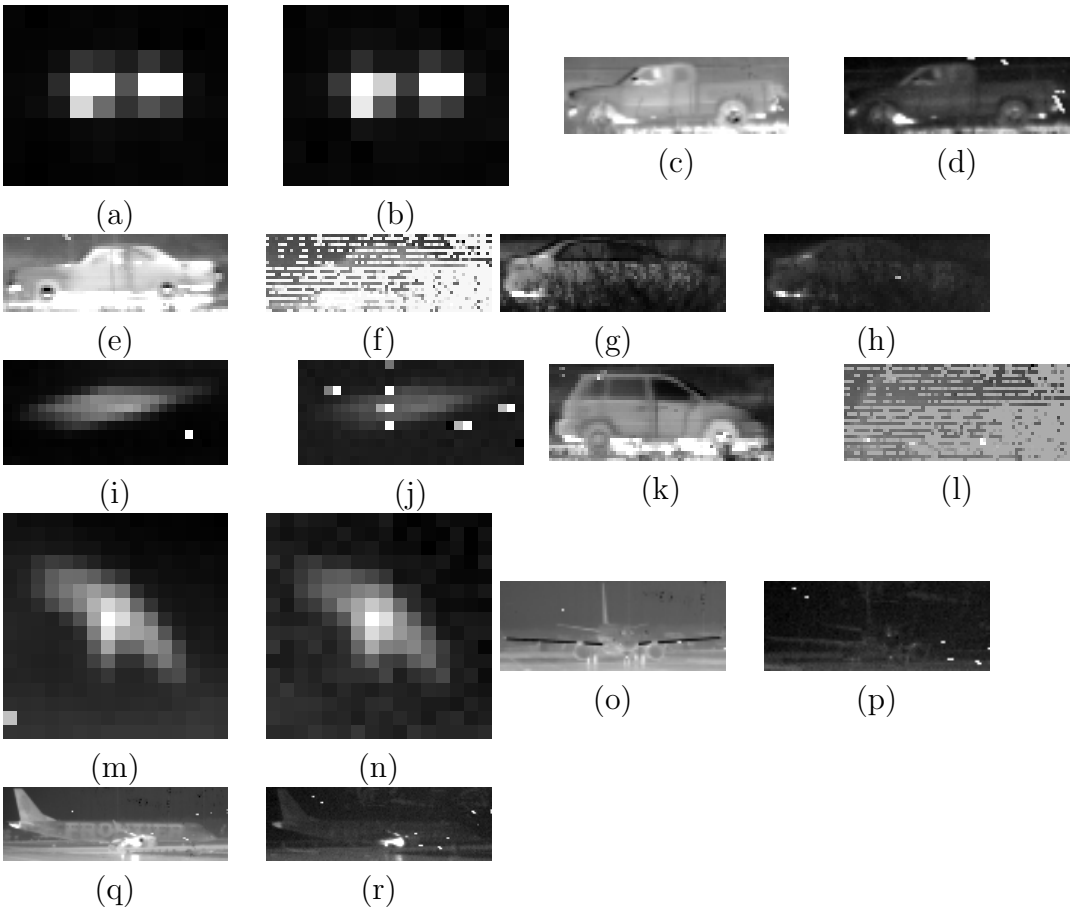


Figure 3.15: Target signatures in sequence SBAP11-SBAP15. (a) SBAP11 LW Object 1; (b) SBAP11 MW Object 1; (c) SBAP12 LW Object 1; (d) SBAP12 MW Object 1; (e) SBAP12 LW Object 2; (f) SBAP12 MW Object 2; (g) SBAP13 LW Object 1; (h) SBAP13 MW Object 1; (i) SBAP13 LW Object 2; (j) SBAP13 MW Object 2; (k) SBAP13 LW Object 3; (l) SBAP13 MW Object 3; (m) SBAP13 LW Object 4; (n) SBAP13 MW Object 4; (o) SBAP14 LW Object 1; (p) SBAP14 MW Object 1; (q) SBAP15 LW Object 1; (r) SBAP15 MW Object 1

SBAP16, SBAP17, and SBAP18 have the same camera setup approximately perpendicular to the roadway. SBAP16 contains three vehicles and a bird. The view SBAP16 is approximately perpendicular to the roadway. Thumbnails for SBAP16, SBAP17 are available in Figs. 3.16 and thumbnails for SBAP18 are available in 3.17. Frame information is available in Tables 3.20.

In the foreground of SBAP16, SBAP17, and SBAP18 there is some oscillatory motion of plant life. This is likely caused by wind driven harmonics of the plants. ViBe's neighborhood learning feature is intended to deal with this type of motion [12]. These sequences would be well suited to evaluate handling this type of motion in motion detection and background extraction algorithms. The oscillatory plant life also partially occludes the target object making these good test cases for detection algorithms.

SBAP19 shows the taxiing and take off of an airplane. The airplane climbs into the sky. Frame information is provided in Table 3.20 and target signature images available in Fig. 3.17(i) and (j).

The initial frames of SBAP20 are of the sky. The camera is then tilted down to show an aircraft taxiing. The first object in SBAP21 is a rear aspect of a somewhat distant aircraft from frames 1 to 81. Then the camera pans to a closer taxiing aircraft and follows the aircraft as it begins accelerating for take off. Frame information is provided for SBAP20 and SBAP21 in Table 3.20. Thumbnails for target signatures are provided in Figs. 3.17(k)-(n) and 3.17(o)-(r) respectively.

SBAP22 is a relatively complex sequence. The camera begins by following Object 2, the aircraft seen in Fig. 3.18 (c) and (d). At some times the

Table 3.20: Objects Instances with Frame Numbers SBAP16-22

Sequence	Object	Start	End	Class
SBAP16	1	1	27	SUV
SBAP16	2	1	52	Car
SBAP16	3	78	129	SUV
SBAP16	4	102	120	Bird
SBAP17	1	126	151	Bird
SBAP17	2	168	218	Car
SBAP17	3	180	301	SUV
SBAP17	4	217	265	SUV
SBAP17	5	304	315	SUV
SBAP17	6	329	374	Pickup
SBAP18	1	1	56	Fuel Truck
SBAP18	2	45	111	SUV
SBAP18	3	77	177	Car
SBAP19	1	1	479	Airplane
SBAP20	1	260	1080	Airplane
SBAP20	2	958	1044	Airplane
SBAP21	1	1	81	Airplane
SBAP21	2	88	1080	Airplane
SBAP22	1	1	7	Car
SBAP22	2	1	640	Airplane
SBAP22	3	62	112	Car

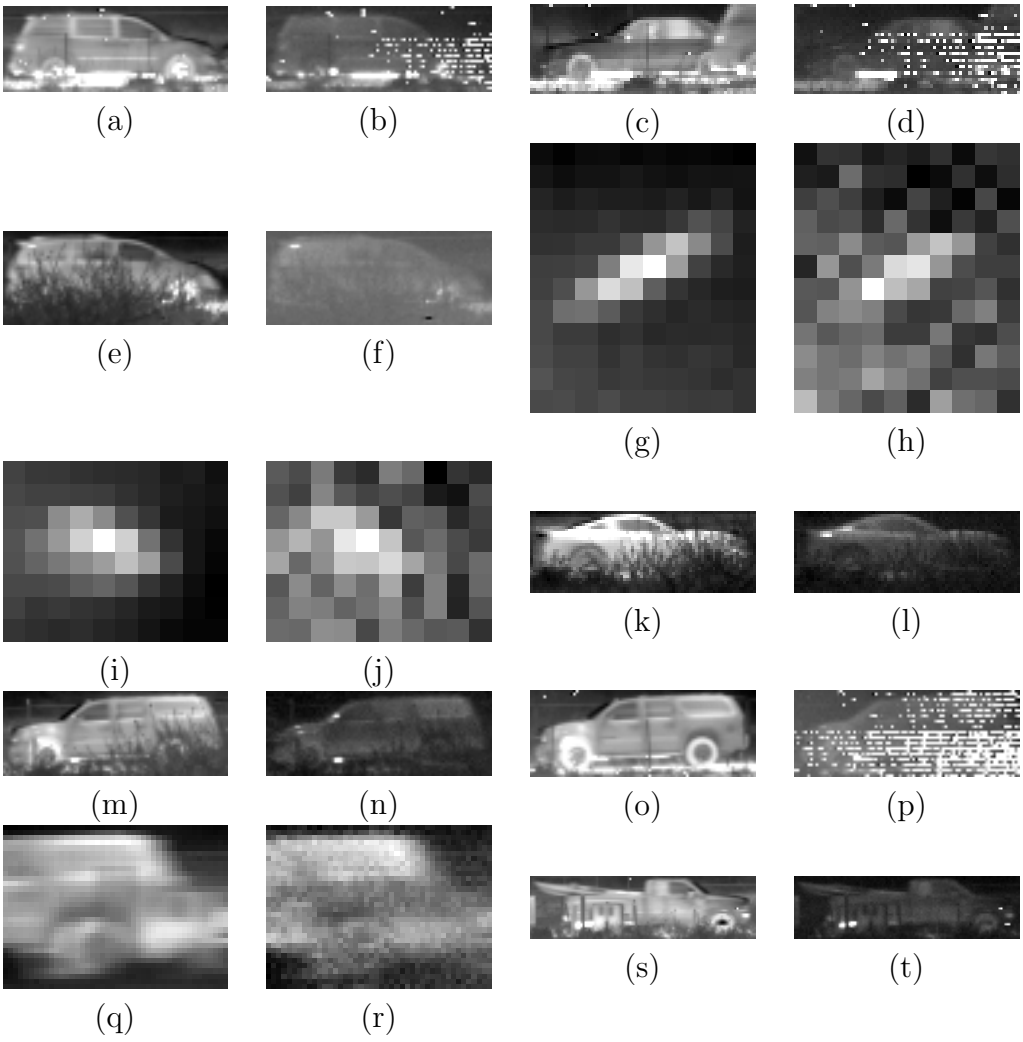


Figure 3.16: Target signatures in sequence SBAP16 and SBAP17. (a) SBAP16 LW Object 1; (b) SBAP16 MW Object 1; (c) SBAP16 LW Object 2; (d) SBAP16 MW Object 2; (e) SBAP16 LW Object 3; (f) SBAP16 MW Object 3; (g) SBAP16 LW Object 4; (h) SBAP16 MW Object 4; (i) SBAP17 LW Object 1; (j) SBAP17 MW Object 1; (k) SBAP17 LW Object 2; (l) SBAP17 MW Object 2; (m) SBAP17 LW Object 3; (n) SBAP17 MW Object 3; (o) SBAP17 LW Object 4; (p) SBAP17 MW Object 4; (q) SBAP17 LW Object 5; (r) SBAP17 MW Object 5; (s) SBAP17 LW Object 6; (t) SBAP17 MW Object 6

camera pans too far in front of the Object 2 and the airplane is out of frame. Corrections were made and the aircraft was reacquired. The panning motion

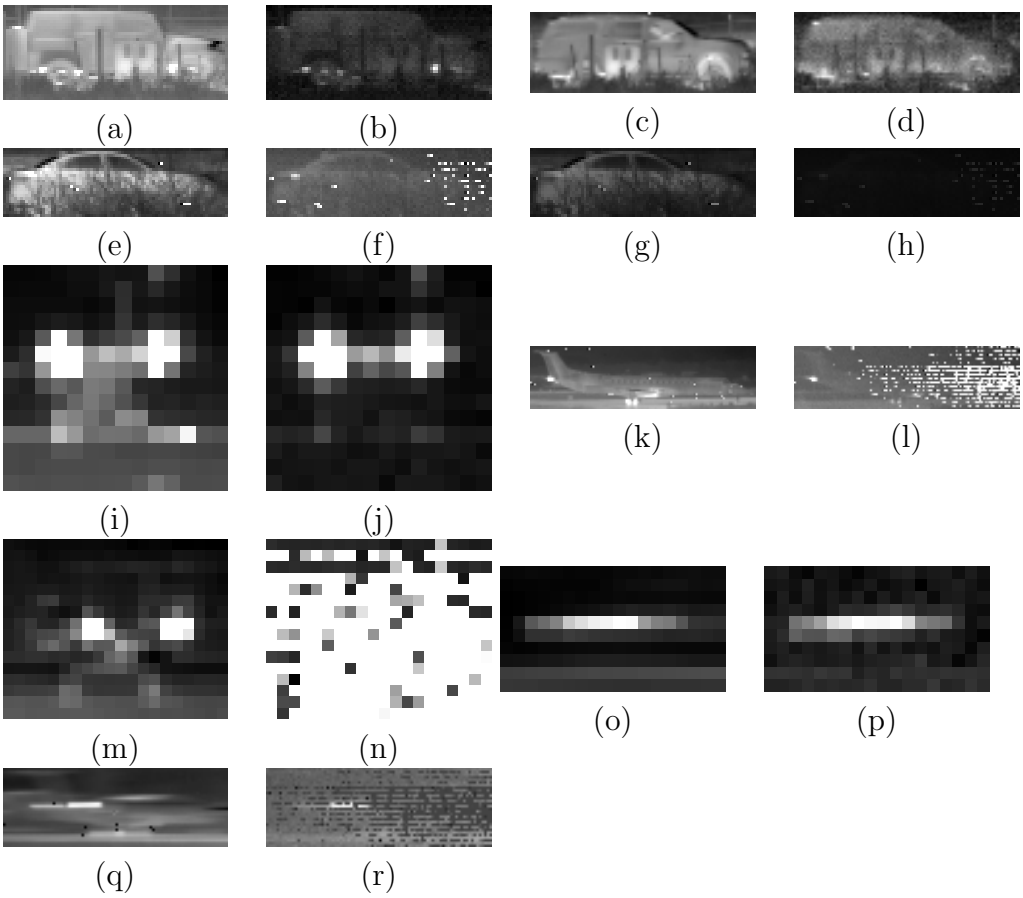


Figure 3.17: Target signatures in sequence SBAP18 - SBAP21. (a) SBAP18 LW Object 1; (b) SBAP18 MW Object 1; (c) SBAP18 LW Object 2; (d) SBAP18 MW Object 2; (e) SBAP18 LW Object 3; (f) SBAP18 MW Object 3; (g) SBAP18 LW Object 4; (h) SBAP18 MW Object 4; (i) SBAP19 LW Object 1; (j) SBAP19 MW Object 1; (k) SBAP20 LW Object 1; (l) SBAP20 MW Object 1; (m) SBAP20 LW Object 2; (n) SBAP20 MW Object 2; (o) SBAP21 LW Object 1; (p) SBAP21 MW Object 1; (1) SBAP21 LW Object 2; (r) SBAP21 MW Object 2

Table 3.21: Objects Instances with Frame Numbers in Sequence SBAP22-24

Sequence	Object	Start	End	Class
SBAP22	1	1	7	Car
SBAP22	2	1	640	Airplane
SBAP22	3	62	112	Car
SBAP23	1	1	410	Airplane
SBAP24	1	1	634	Fuel Truck

captures creates a dynamic background image with the hills in the background. Two vehicles were captured on the roadway as the aircraft passed overhead. Frame information is provided in Table 3.20.

SBAP23 is similar to SBAP22 in that it follows an aircraft on approach. However, in this case the camera is slightly underpanned and the aircraft passes out of frame. No ground vehicles were captured in SBAP23, frame information is in Table 3.21 and thumbails in Fig. 3.18(g) and (h).

SBAP24 captures a fuel truck as it crosses in front of the camera from left to right and the makes a turn following the perimeter road. The appearance of the truck changes significantly as its aspect changes through the turn. Target signatures are in Fig. 3.18(i) and (j) and frame information is provided in Table 3.21.

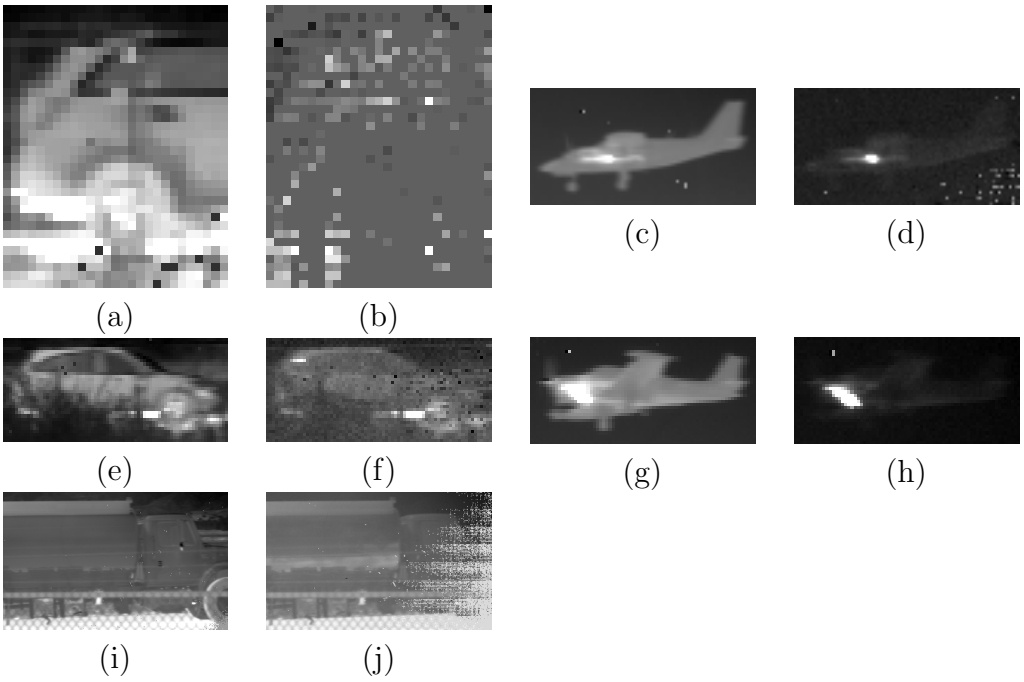


Figure 3.18: Target signatures in sequence SBAP22 - SBAP24. (a) SBAP22 LW Object 1; (b) SBAP22 MW Object 1; (c) SBAP22 LW Object 2; (d) SBAP22 MW Object 2; (e) SBAP22 LW Object 3; (f) SBAP22 MW Object 3; (g) SBAP23 LW Object 1; (h) SBAP23 MW Object 1; (i) SBAP24 LW Object 1; (j) SBAP24 MW Object 1

Similar to sequences SBAP22 and SBAP23, in SBAP25 the camera follows an aircraft in the landing process. This sequence more successfully captures the landing sequence only briefly losing the airplane. The airplane is represented as two sequences Object 2 and Object 3 as it was off the screen. Two vehicles were also captured in this sequence. Significant ego motion in the form of panning an tilt adjustment were change the background of the sequence. Target signatures are in Fig. 3.19 and frame information is provided in Table 3.22. SBAP26, similar to SBAP25, follows an aircraft on approach. This sequence starts later in the descent and manages to capture the aircraft landing and decelerating on the runway. Target signatures are in Fig. 3.19(i)

Table 3.22: Objects Instances with Frame Numbers in Sequence SBAP25-28

Sequence	Object	Start	End	Class
SBAP25	1	1	37	Car
SBAP25	2	1	515	Airplane
SBAP25	3	552	640	Airplane
SBAP25	4	344	368	SUV
SBAP26	1	1	639	Airplane
SBAP27	1	77	696	Fuel Truck
SBAP27	2	781	1279	Airplane
SBAP28	1	1	1280	Airplane

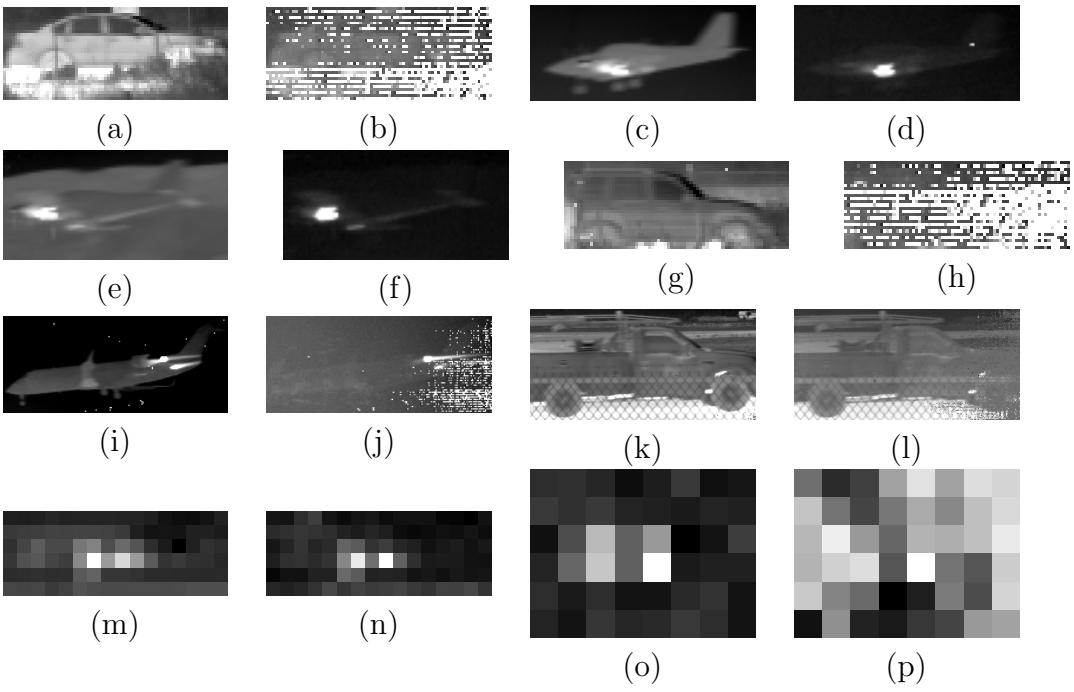


Figure 3.19: Target signatures in sequence SBAP25-28. (a) SBAP25 LW Object 1; (b) SBAP25 MW Object 1; (c) SBAP25 LW Object 2; (d) SBAP25 MW Object 2; (e) SBAP25 LW Object 3; (f) SBAP25 MW Object 3; (g) SBAP25 LW Object 4; (h) SBAP25 MW Object 4; (i) SBAP26 LW Object 1; (j) SBAP26 MW Object 1; (k) SBAP27 LW Object 1; (l) SBAP27 MW Object 1; (m) SBAP27 LW Object 2; (n) SBAP27 MW Object 2; (o) SBAP28 LW Object 1; (p) SBAP28 MW Object 1

and (j) and frame information is provided in Table 3.22. The appearance of the aircraft changes significant through the course of this sequence. The aircraft is seen nearly in profile at the beginning of the sequence the as the aircraft lands and travels into the distance a view of the jet engines from behind is apparent.

SBAP27 is a sequence following a passing utility pickup on the perimeter road. Target signatures are in Fig. 3.19(j)-(n) and frame information is provided in Table 3.22. The sequence displays significant ego-motion and after the Fuel Truck passes on the perimeter road the camera pans to a taxiing aircraft.

SBAP28 is a particularly challenging sequence to label. There is a distant jet engine signature that was challenging to see without employing visual enhancements. Thumbnails representing the largest appearance of the signature presented in Fig. 3.19(o) and (p) and frames in Table 3.22.

3.0.3 Vons Grocery, Bishop, CA Sequences

Six sequences were captured in the parking lot at Vons Grocery Store in Bishop, CA. The sequences are all captured at night and capture activity in the parking lot. There are several people and cars moving. Some vehicles are stationary in parking spaces for the duration of the sequence.

Tables describing the start and stop frames, object number, and class are provided for each sequence captured at Von's in Tables. 3.23, 3.24, 3.25, 3.26, 3.27, and 3.28. Thumbnails of these sequences are provided in Figs. 3.20, 3.21, 3.22, 3.23, 3.24, and 3.25. Vons1 objects 3,4, and 5 are people moving near to the camera and there is significant blur to the objects. Objects 1 and 2 undergo several significant occlusions. All objects but one in vons6 are people moving, which is a parked car. This is a challenging sequence of tracks

Table 3.23: Objects Instances with Frame Numbers in Sequence vons1

Sequence	Object	Start	End	Class
vons1	1	1	1070	Person
vons1	2	1420	1657	Person
vons1	3	2058	2080	Person
vons1	4	2063	2065	Person
vons1	5	2339	2341	Person
vons1	6	2339	2339	Car
vons1	7	2376	2379	Pickup

for people with several occlusion events, and people leaving and re-entering the camera frame.

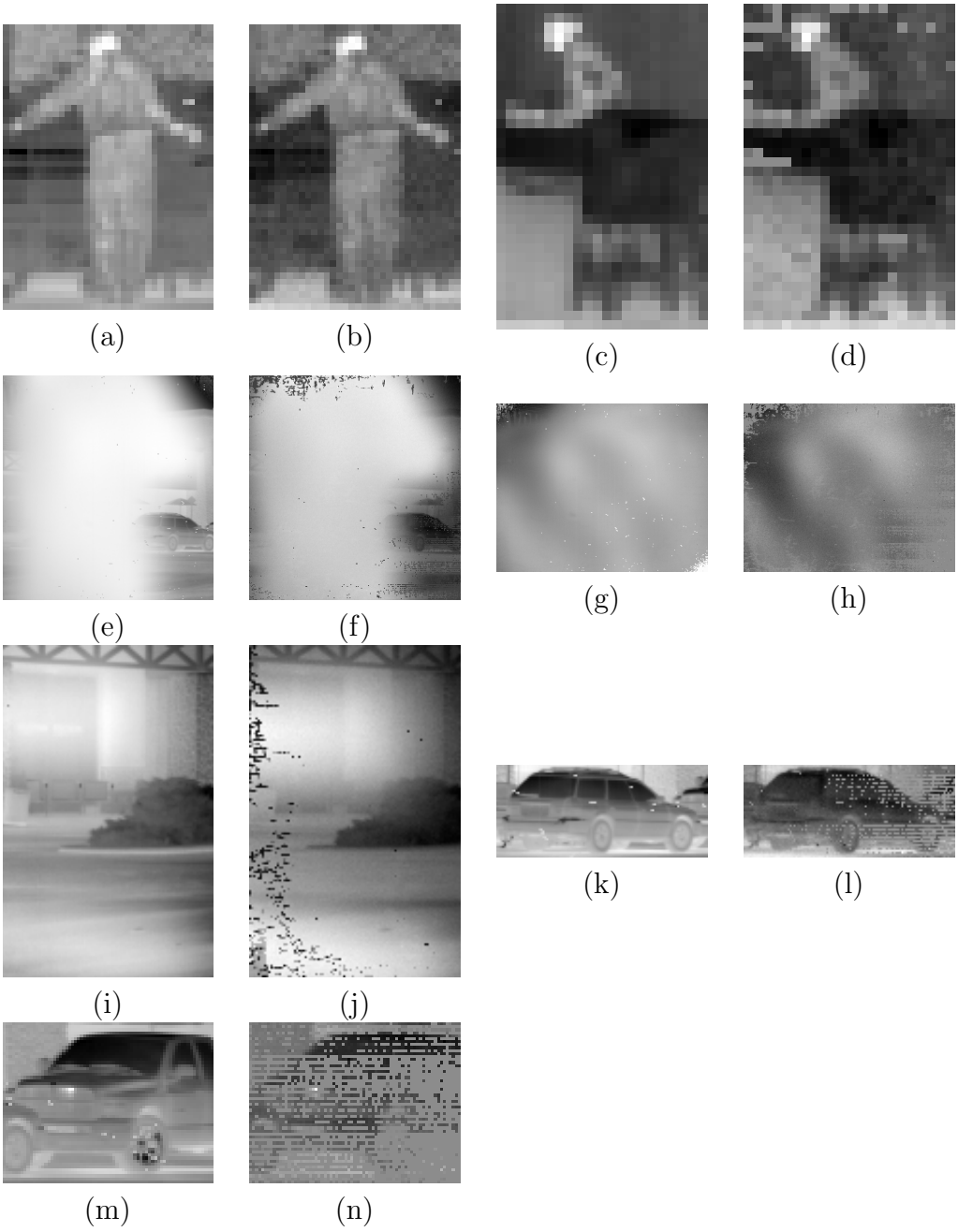


Figure 3.20: Target signatures in sequence vons1. (a) LW Object 1; (b) MW Object 1; (c) LW Object 2; (d) MW Object 2; (e) LW Object 3; (f) MW Object 3; (g) LW Object 4; (h) MW Object 4; (i) LW Object 5; (j) MW Object 5; (k) LW Object 6; (l) MW Object 6; (m) LW Object 7; (n) MW Object 7

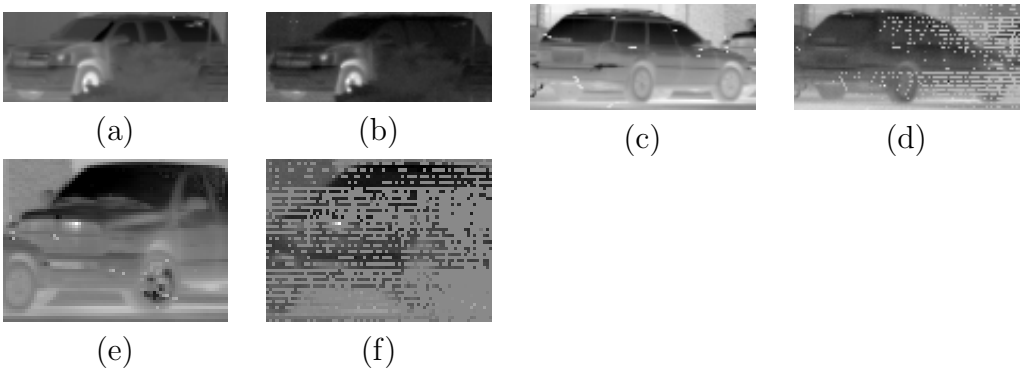


Figure 3.21: Target signatures in sequence vons2. (a) LW Track 1; (b) MW Track 1; (c) LW Track 2; (d) MW Track 2; (e) LW Track 3; (f) MW Track 3

Table 3.24: Objects Instances with Frame Numbers in Sequence vons2

Sequence	Object	Start	End	Class
vons2	1	1060	1200	SUV
vons2	2	1	1200	Car
vons2	3	1	1200	Pickup

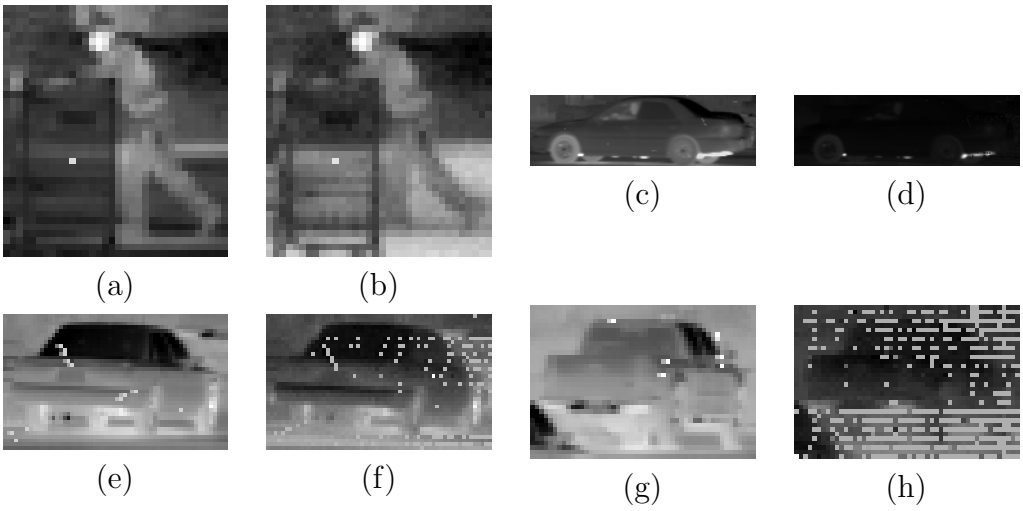


Figure 3.22: Target signatures in sequence vons3. (a) LW Track 1; (b) MW Track 1; (c) LW Track 2; (d) MW Track 2; (e) LW Track 3; (f) MW Track 3; (g) LW Track 4; (h) MW Track 4

Table 3.25: Objects Instances with Frame Numbers in Sequence vons3

Sequence	Object	Start	End	Class
vons3	1	1	211	Person
vons3	2	707	765	Car
vons3	3	1	1200	Car
vons3	4	1	1200	Pickup

Table 3.26: Objects Instances with Frame Numbers in Sequence vons4

Sequence	Object	Start	End	Class
vons4	1	1	609	Person
vons4	2	669	700	Person
vons4	3	804	968	Person
vons4	4	1	607	Pickup
vons4	5	1	572	Car
vons4	6	561	838	Pickup
vons4	7	574	1199	Car

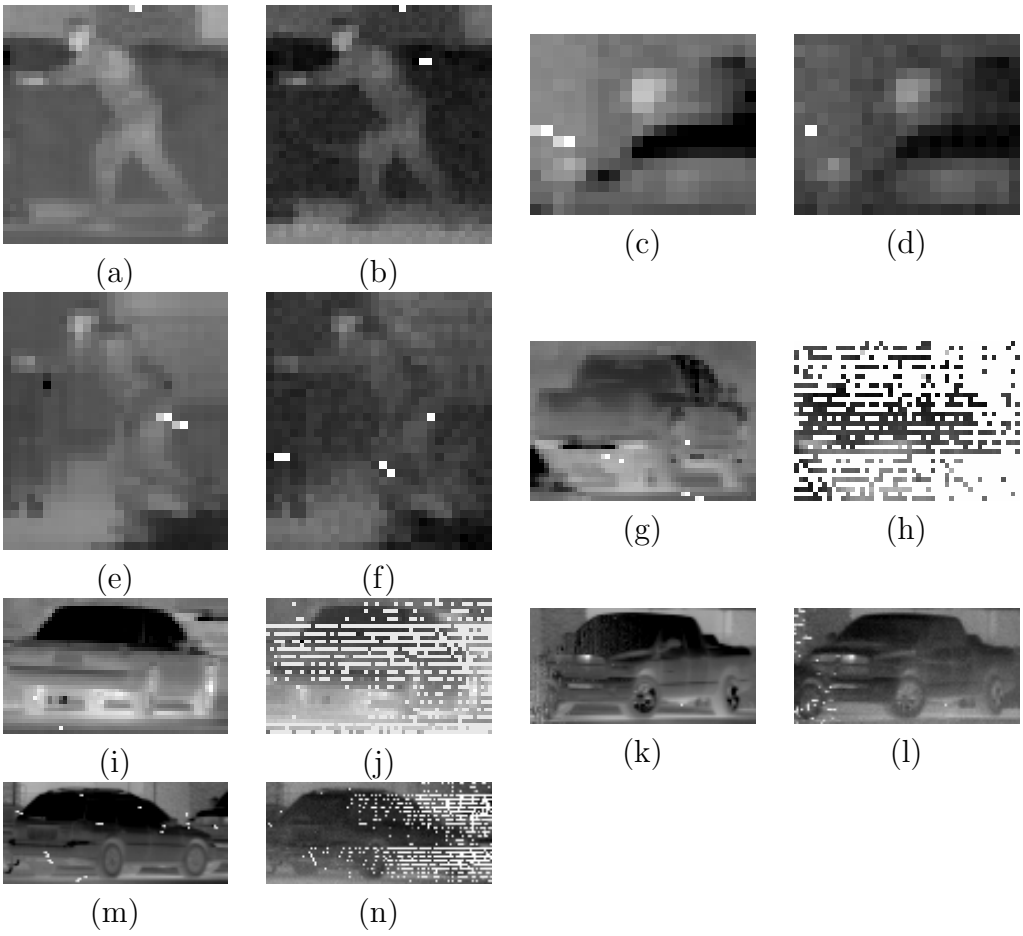


Figure 3.23: Target signatures in sequence vons4. (a) LW Track 1; (b) MW Track 1; (c) LW Track 2; (d) MW Track 2; (e) LW Track 3; (f) MW Track 3; (g) LW Track 4; (h) MW Track 4; (i) LW Track 5; (j) MW Track 5; (k) LW Track 6; (l) MW Track 6; (m) LW Track 7; (n) MW Track 7

Table 3.27: Objects Instances with Frame Numbers in Sequence vons5

Sequence	Object	Start	End	Class
vons5	1	1	228	Person
vons5	2	407	906	SUV
vons5	3	1517	1800	SUV
vons5	4	1	1800	SUV

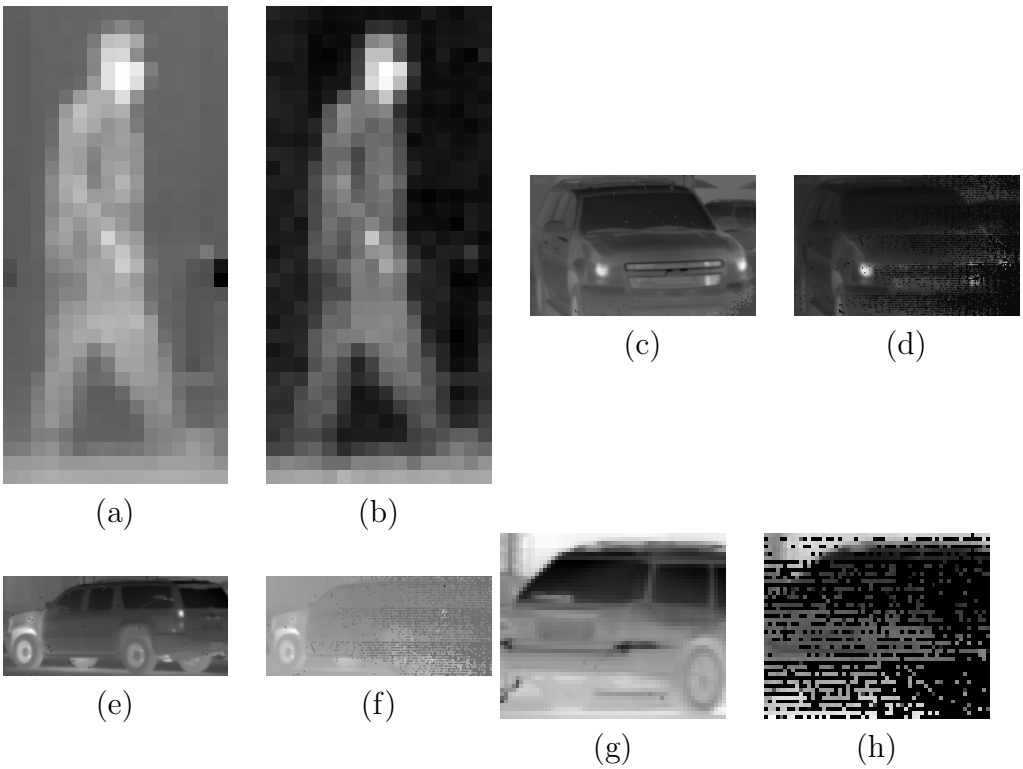


Figure 3.24: Target signatures in sequence vons5. (a) LW Track 1; (b) MW Track 1; (c) LW Track 2; (d) MW Track 2; (e) LW Track 3; (f) MW Track 3; (g) LW Track 4; (h) MW Track 4

Table 3.28: Objects Instances with Frame Numbers in Sequence vons6

Sequence	Object	Start	End	Class
vons6	1	1	1092	Person
vons6	2	1	1300	Person
vons6	3	1	235	Person
vons6	4	1323	1548	Person
vons6	5	1338	1391	Person
vons6	6	1415	1445	Person
vons6	7	1515	1578	Person
vons6	8	1	1800	Person

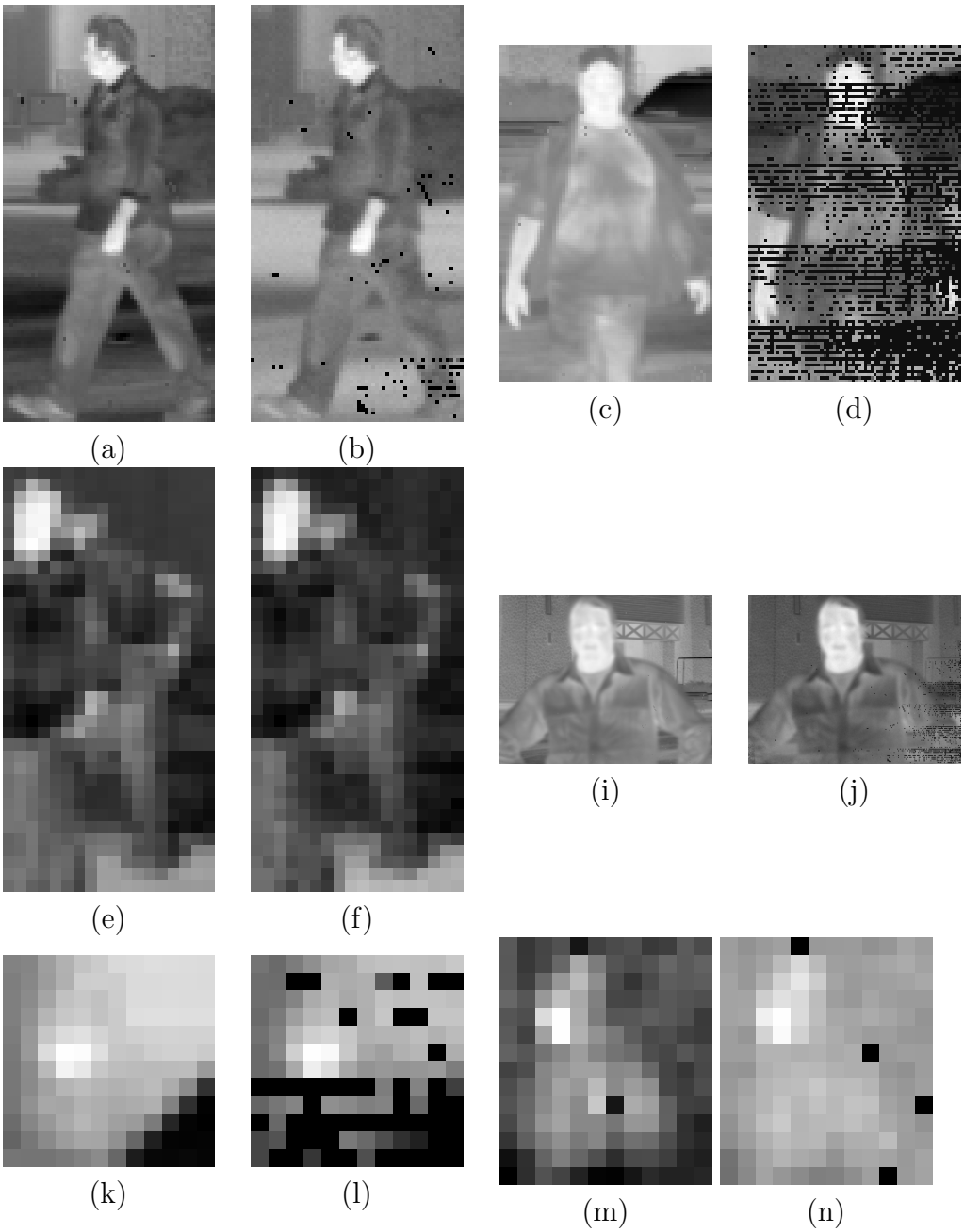
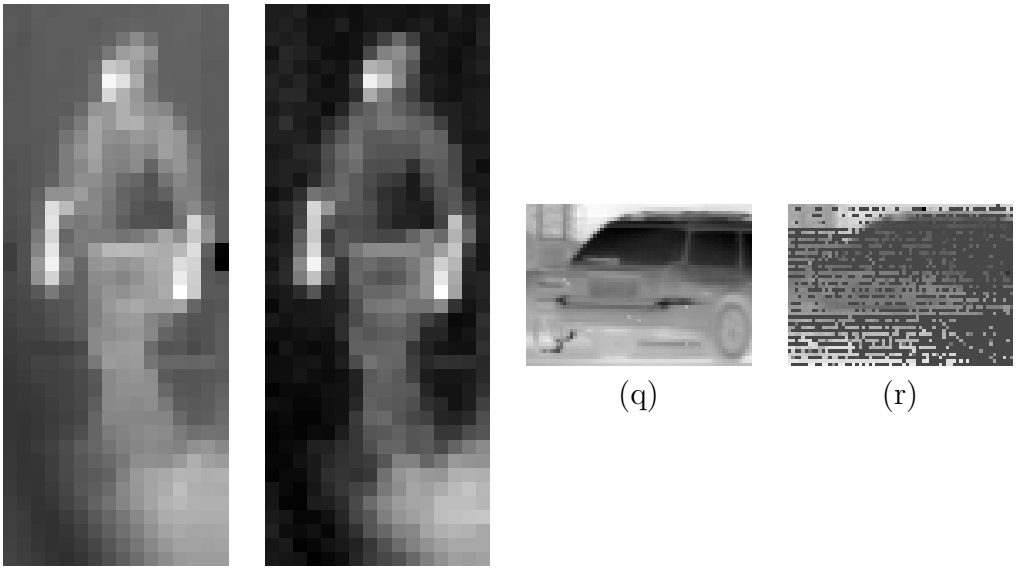


Figure 3.25: Target signatures in sequence vons6. (a) LW Track 1; (b) MW Track 1; (c) LW Track 2; (d) MW Track 2; (e) LW Track 3; (f) MW Track 3; (g) LW Track 4; (h) MW Track 4; (i) LW Track 5; (j) MW Track 5; (k) LW Track 6; (l) MW Track 6; (m) LW Track 7; (n) MW Track 7



(o) (p) (q) (r)

Figure 3.25: Target signatures in sequence vons6. (o) LW Track 8; (p) MW Track 8

There are at least 15,008 frames with one object obscuring the view of another object.

The SQL query provided in Appendix B can be used to calculate IoU in a MySQL DB. The query as written compares ground truth to ground truth. If there is overlap of ground truth then one Object is obscuring another. Modifications to this script, say changing one of the "ground truth" source tables to a "detection table" with the proper formatting can be used to evaluate detection performance.

I designed the SQL script in Appendix B was used to create a table of IoUs for each detection time. For ViBe IoU there was also a field that included the spectrum of the detection MW, LW, or DB. For YOLOv4 additional columns for spectrum and thresholded were added to keep track if the detection was MW or LW and if the image had been preprocessed. The number of labeled objects in the dataset is detailed in Table 3.29.

Table 3.29: Object Class Count

Class	Count
Airplane	23
Bird	16
Box Truck	2
Car	63
Fuel Truck	5
Indeterminate	14
Person	19
Pickup	44
Pickup and Trailer	3
Pickup?	1
Semi	1
Semi and Trailer	4
SUV	29
SUV and Trailer	4
Van	4

In this chapter, a large diverse DBIR data set with high utility for object detection and tracking was introduced. While this dataset may not be sufficiently large to significantly retrain a CNN, as the 233 objects likely do not present broad enough variety of appearance to provide generality and to avoid overfitting. There are a substantial number of long track objects which makes this data set excellent for evaluating target tracking algorithms.

Chapter 4

Experiments

4.1 Experimental Setup

To better understand the benefits of DBIR over LWIR or MWIR alone in the context of a deployable modern computer vision detection system, four algorithms were selected for evaluation. The CNN based YOLOv4 and YOLOv7 were chosen because they represent the state of the art for real-time deep learning based detectors. ViBe was selected as it represents one of the best performing motion detection algorithms. Additionally, a simple threshold-based detector was developed for evaluation. Subsequently, all four of these algorithms were assessed separately for each spectrum using the images described in Chapter 3.

4.1.1 CNN Experiment Setup

A significant amount of pre-processing, as described below, was required to achieve desirable performance from the CNNs. After pre-processing, the MWIR images and LWIR images were evaluated separately using instances of YOLOv4 and YOLOv7, each trained on MS COCO. The detections for each frame of the data set from each CNN were collected into a database. Subsequently, the IoU for each detection was calculated against each ground truth per frame. To label a detection as a TP the maximum IoU for that detection, relative to all ground truth objects in that frame, had to exceed a threshold. Otherwise, the

detection was labeled a False Positive (FP). If the maximum IoU for a ground truth object within a frame failed to exceed that threshold, it was considered a False Negative (FN). DBIR in this context meant considering detections form both LWIR and MWIR together. Analysis was conducted on these results to understand the sensitivity of the object detectors improvement using DBIR in terms of object size, IoU Threshold, and Confidence thresholding.

4.1.2 Motion Detection Based Experiment Setup

To evaluate a modern motion detection algorithm and better understand the benefits of DBIR over MWIR or LWIR alone, I implemented the parallel version of ViBE, as described in Chapter 2, MATLAB. Each frame of every labeled sequence was processed by the motion detection algorithm. The number of samples for each pixel was set to 20, and the first 20 frames were used as the initial background model. Each pixel of every frame was labeled as foreground/background which I call a 'foreground map.'

To evaluate DBIR in this context, the foreground maps for LWIR and MWIR were combined by pixel-wise logical OR operation. In other words, if the pixel was foreground in either LWIR or MWIR, it was considered as foreground in DBIR. Morphological operations were then applied to the three foreground maps to remove small isolated foreground and background pixels. The morphological operations are described in detail in the algorithm description in Section 2.6.1.

After morphological operations, each foreground map was labeled using Connected Components labeling. Each connected component was turned into a bounding box by taking the maximum and minimum pixel location in each

dimension of the connected component. This process provided the coordinates of the four corners of the bounding box in the two-dimensional pixel plane. These detections were then used to create a database for ViBE detection in MWIR, LWIR, and DBIR.

Database detection for ViBe and YOLO were labeled as TP or FP using the same process. A script has been provided in Appendix B with details of the IoU generation.

At the pixel, I conducted an experiment involving thresholding bright spots. This experiment proved effective at detecting jet exhausts in some Santa-Barbara airport video sequences. However, there was very little benefit of DBIR over MWIR or LWIR alone. In my experience, it is unlikely that this approach would improve detection. The target signatures for jet engine exhaust plumes were very similar in MWIR and LWIR, as illustrated in Fig. 3.17(i),(j),(m)-(p) and Fig. 3.19(m)-(p).

Sect. 4.1.5 briefly discusses how an engineer could integrate these distinct detection systems into a coherent framework and highlights the potential benefits stemming from integrating these two detection methods.

4.1.3 Dual-Band Motion Detection with ViBe

The parallelized version of ViBe, described in Section 2.6.1, was implemented in MATLAB for the purposes of conducting these experiments. The background model was initialized with the first N frames, with N set to 20 for all experiments. Subsequently, each frame of the sequence was processed through the algorithm.

In Fig. 4.1, the process of comparing the incoming image to the back-

ground model is illustrated. Most of the processing of ViBe occurs at the pixel level. In the original implementation in [12], the pixels are processed in raster scan order. However, in my implementation, pixels are processed in parallel. Determining if the pixel of an incoming image is foreground involves checking the disparity between an incoming pixel value and any of the background samples at that pixel location. If the disparity is less that a given value, R, a count of background values the incoming pixel is "close to" is incremented. When the count exceeds another threshold, using the notation from [12] $\#_{min}$, that pixel is considered a background value. For all reported experimental results here, the parameter R, indicating the distance between the incoming pixel and background value is be considered sufficiently close, was set to 90. $\#_{min}$ was set to 4 or 20% of the background model samples for that pixel location. ViBe learns background samples from incoming images to attempt to adapt to condition changes of the background. The learning rate parameter, ϕ, was set to 1/8 for these experiments.

In tandem to ViBe, I implemented a object detection method similar to the one describe in [112]. Morphological operations, namely a morphological CLOSE operation, connect clusters of foreground detections near each other and remove small internal concavities in the Raw Foreground map. It has the additional effect of removing small isolated pixels caused by scintillation and speckling. This new image undergoes a connected components analysis. Connected components labels pixels adjacent to similar pixels with the same label. An example of the connected components output can be seen in fig. 4.1. The dark blue is the background and each connected component is a different color. Then for each connected component a bounding box is generated. The

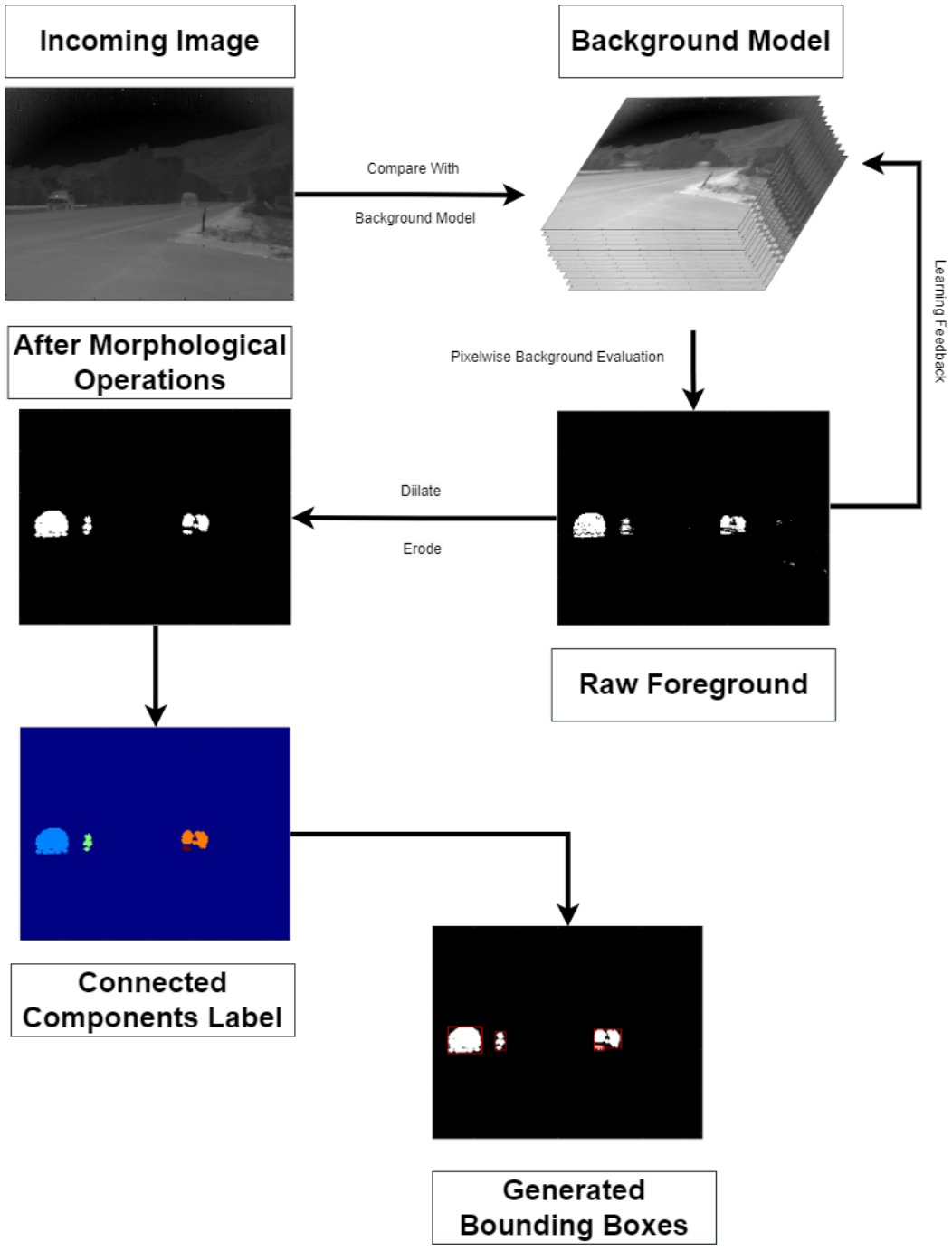


Figure 4.1: Diagram of ViBe Detection Process.

bounding box is what we consider the "detection" for our purposes and is comparable to the bounding box output of a CNN like YOLO or RCNN. This is not my original idea and other papers have used this approach [199].

To visualize the learning process the diagram, in Fig. 4.2 shows the process of removing the foreground pixels from the randomly generated learning locations to remove the learning of foreground from the process as is described in [12]. For our algorithm, we also randomly learn foreground pixels. This process is called "eaten up" and was suggested in [179]. A pixel-wise performance analysis of the improvement of adding "eaten-up" to ViBE was not evaluated but it had a strong impact qualitatively. In particular, it ameliorated the "waking ghost" problem which is the phenomenon of having a stationary object in the video sequence that starts moving after it has been incorporated into the background model. In that situation the space where the object was remains a foreground detection if the background model never learns foreground. Getting the learning parameter properly tuned is of central importance to the success of "eaten up."

If the object is moving quickly enough relative to the learning process and especially relative to $\#_{min}$ there is little risk of the object being learned as background. By randomly learning the foreground, waking ghosts eventually erode and the background is learned.

The values of the incoming image at the learning locations is inserted into one of the sample locations in the background model. This gives the background model a slow but consistent learning process. Motion detection models are often extremely sensitive to the motion of the camera itself. Such motion often induces a significant portion of the image to be detected as foreground.

This continues until the camera has become stationary and the background model converges on the new scene. Fig. 4.3 is a diagram with a visual explanation of the data fusion process.

Combining the information between MWIR and LWIR bands of this data set represents a novel contribution. The approach taken was to combine the information at the Raw Foreground stage, shown in Fig. 4.1. The data fusion technique used was to label a pixel as DB foreground if it was identified as foreground in either the MW or LW spectrum. Subsequently, for evaluation bounding boxes were generated for MW, LW, and DB separately. These bounding boxes can then be compared to the ground truth bounding boxes to calculate the IoU.

Table 4.1 and Table 4.2 show the number of objects in sequences without significant ego-motion of the camera. In the left hand column is the sequence name. In DB $>$ MW are the number of detections in which the DB detection achieved a higher IoU than the MW detection of the same object. In the column DB $<$ MW the number of detections in which the MW had a higher IoU than the DB. Then percent improved is show as a fraction of the total and percent deteriorated is also shows. The bottom row give totals and over all percentages. Overall DBIR improved MWIR detections 81.64% of the time at a cost of of 8.92%. Table 4.1 is structurally similar to Table 4.2 but looks at LW instead of MW. In comparison LW was improved by adding DBIR data 7.81% at a cost of 1.70%.

As mentioned in Chapter 2, most background estimation techniques for motion detection look at pixel-wise segmentation rather than bounding boxes. A subset of the available sequences were used to evaluate the performance of the

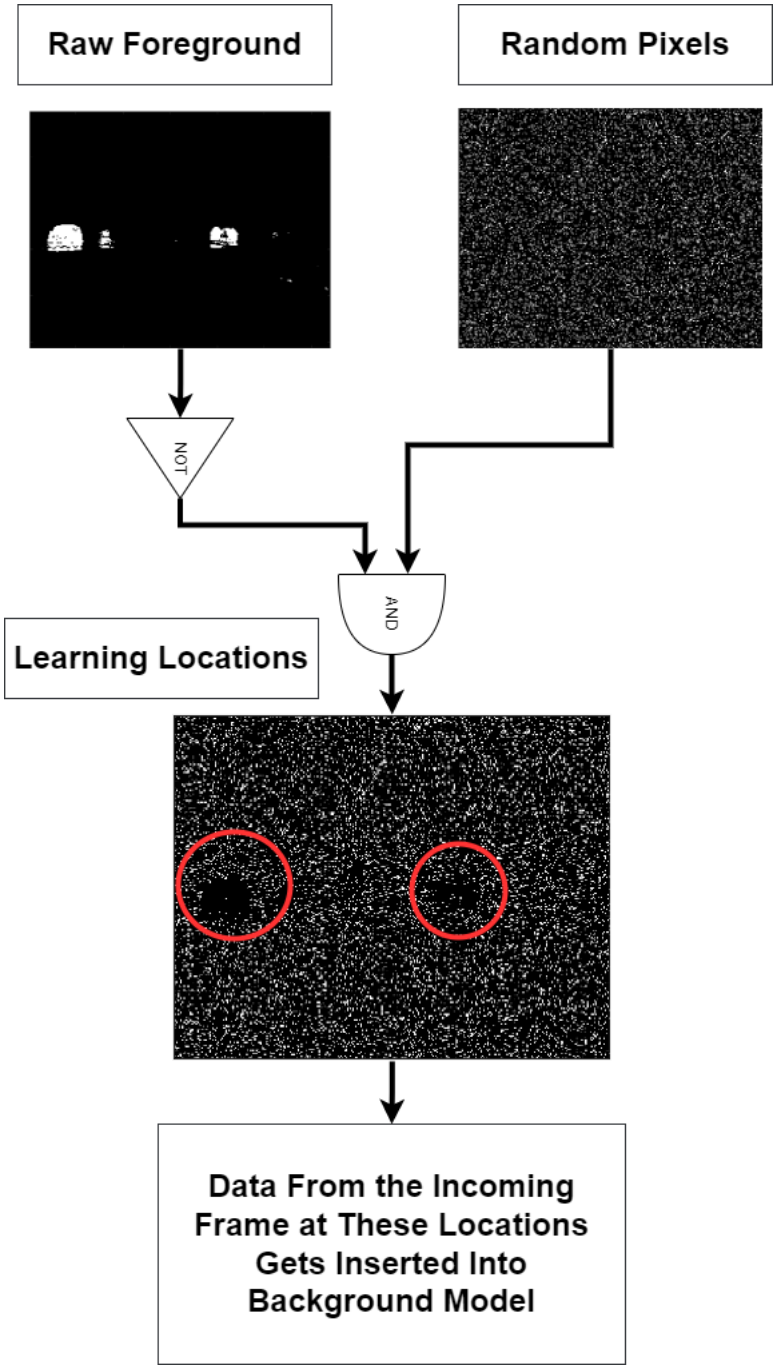


Figure 4.2: ViBe Learning Location Selection Process. Red circles the removal of learning locations in foreground areas.

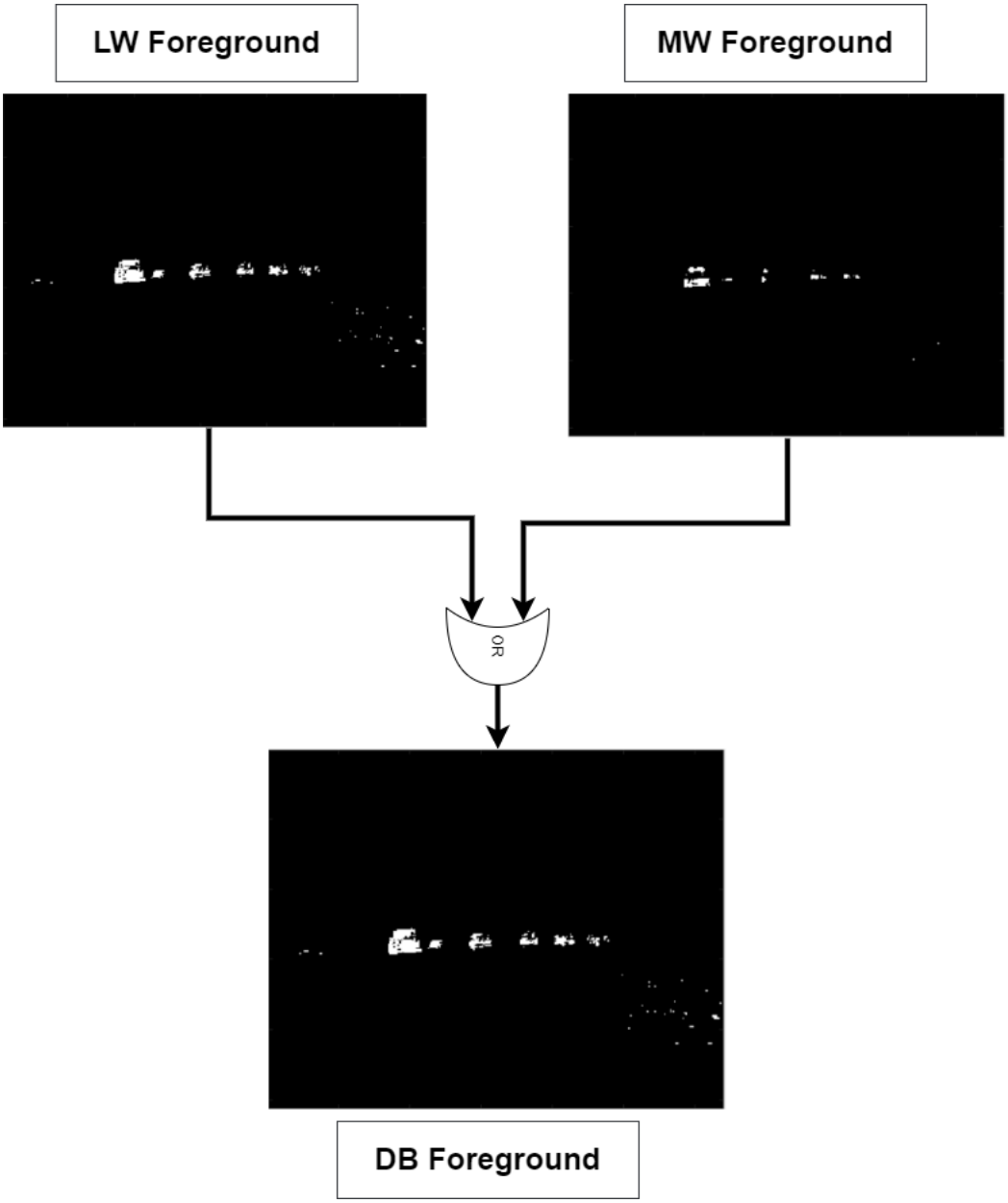


Figure 4.3: Pixel-wise OR-ing Operation for motion detection fusion.

object detection via motion. The sequences with extended stationary camera views can be seen in the left hand column of Table 4.1 and Table 4.2.

Table 4.1 shows the benefits of DB data over LW alone for these se-

quences. For each ground truth object in the sequences selected the maximum IoU for that object was determined. Then the maximal IoUs for each ground truth object were compared per spectrum MW, LW, and DB. The total count for when DB exceeds LW or MW alone, for when LW orMW exceed DB and when they are the same are shown in the second through fourth columns of Tab. 4.1 and 4.2.

The primary reason for the benefits of DBIR over separate LWIR and MWIR is the variation in the emissivity of materials in different parts of the spectrum. Under certain circumstances within the data set, parts of the object that are emissive in the MW are less emissive in LW and vice versa. Given that regions of the object appear in different parts of the spectrum, fusing the data caused the foreground regions in each spectrum to become connected. This subsequently leads to bounding boxes in the connected components stage to more closely match the bounding box as represented in the ground truth.

This phenomenon is apparent in Fig. 4.4, where the person's mid-section is detected more completely as foreground in the MW while the legs arms are more completely pixel-wise detected in LW. Combining these two creates a more complete understanding of the object's size. Fig. 4.4 contains MW ViBe detections on the top row, LW ViBe detections in the middle row, and DBIR ViBe detections in the bottom row. The left hand column are detection bounding boxes over raw images. The right hand column of this figure contains the same bounding boxes over foreground maps. The side by side comparison is presented to give the reader an understanding of the object that is being detected and the internal foreground map state of the detector.

At times the IoU decreases with ViBe. Per observation, this is almost

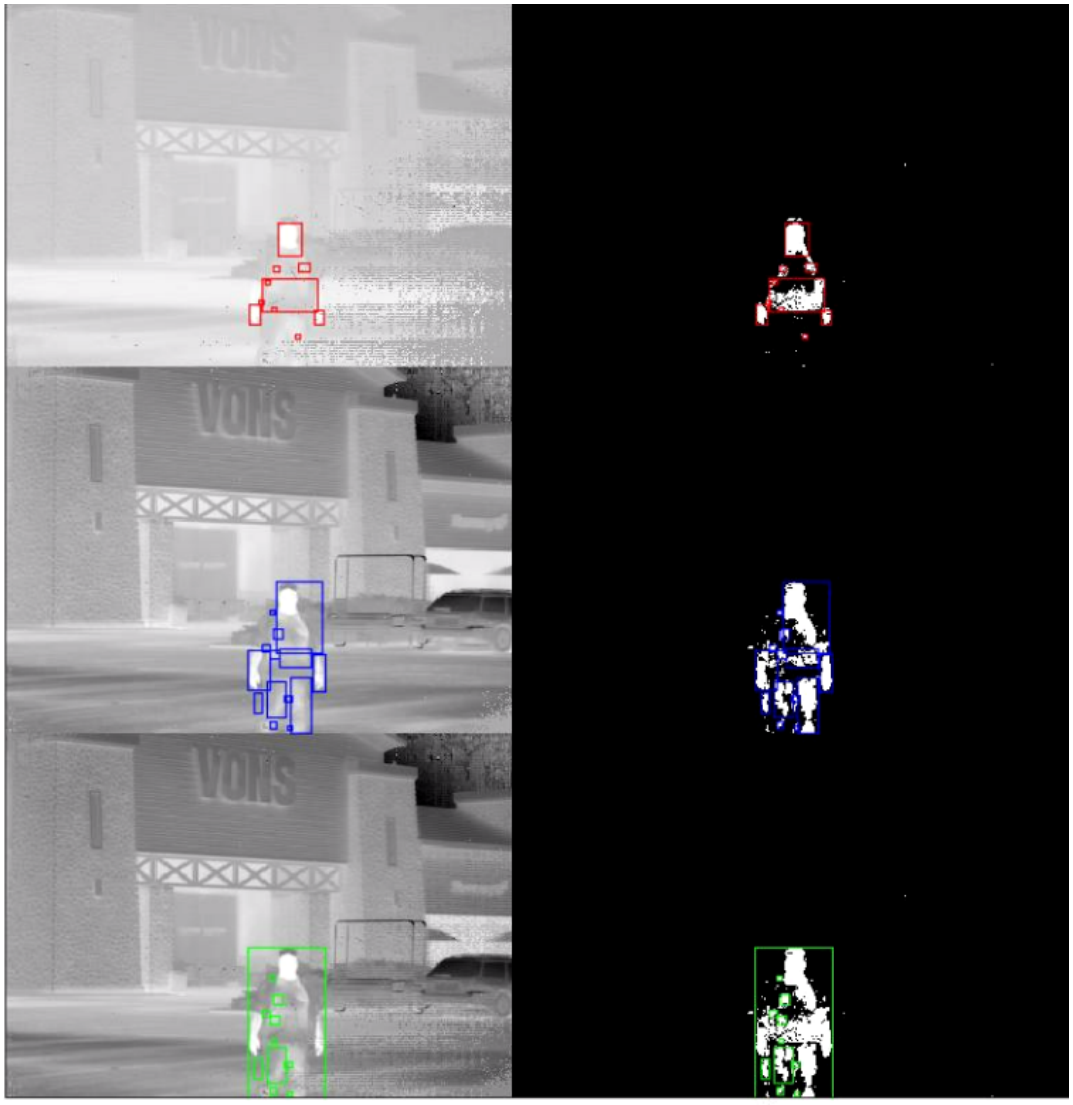


Figure 4.4: ViBe Raw Image with Bounding Boxes and Motion Detection with Bounding Boxes. This image shows the biggest improvement in the experiment between LW and DB and is representative of the nature of the performance improvement. Top row is MW, middle row is LW, and bottom row is DB.

exclusively caused by one ground truth object occluding another ground truth object. If this happens the connected components operations of the algorithm presented here will combine them. There is a potential that the maximal IoU for both objects will be deteriorated by this combination in the DB estimate.

This will count against DB more than once in Table 4.1 and 4.2 because each of the ground truth objects IoU is effected by the merging. An instance of this occlusion phenomenon is illustrated in Fig. 4.5. In Fig. 4.5, similar to Fig. 4.4, raw images are seen on the left and foreground maps are seen on the right. However, unlike in Fig. 4.5 where separate parts of an object get merged two separate objects gets merged. This creates a false positive with the overly large bounding box and two missed detections. One approach to minimizing the detection merging problem, would be to use a multi-object tracker which would track the bounding boxes as they approach each other. The occlusion condition could then be predicted and as tracks associate underlying detected pixels they could in a way uncluster the pixelwise detections. This algorithm is outside of the scope of this document and represents a potentially beneficial avenue of research.

Another thing that can be observed from Table 4.1 and 4.2 is that MW's coverage on average is much less the LW. That is, MW is either more segmented in its coverage. The more segmented having lower scores is because we are looking at maximal IoU over what may be many bounding boxes over the object.

It is evident from Table 4.1 that, on net, 6.11% of LW detections were improved by adding information from the MW. For MW, the addition of LW information improved 72.72% of detections. If pixel-wise segmentation was in the budget, it would be interesting to determine the per pixel differences. Motion detection based foreground/background detection becomes significantly more difficult when there is a camera motion; while sophisticated techniques for estimating and compensating for camera motion have been investigated [2,

Figure 4.5: ViBe Raw Image with Bounding Boxes and Motion Detection with Bounding Boxes. This figure shows the biggest deterioration of IoU from MW and LW to DB. The cause of the problem is that two objects have been merged. Top row is MW, middle row is LW, and bottom row is DB.

157, 180], the results of motion detection are often unreliable in the presence of significant camera motion.

An interesting observation stems from the distinct detection distances

Table 4.1: Comparison of DB IoU to LW IoU

Sequence	DB > LW	DB < LW	DB = LW	%-imp.	%-det.
brwncamp1	413	22	8,747	4.50	0.24
brwncamp2	467	79	5,281	8.01	1.36
brwncamp3	189	67	4,376	4.08	1.45
brwncamp5	440	175	6,192	6.46	2.57
brwncamp6	107	96	5,508	1.87	1.68
brwncamp7	697	225	3,207	16.88	5.45
brwncamp8	867	195	8,118	9.44	2.12
SBAP12	15	1	94	13.64	0.91
SBAP13	8	0	146	5.19	0.00
vons1	31	8	1,271	2.37	0.61
vons2	1	0	137	0.72	0.00
vons3	11	1	236	4.44	0.40
vons4	12	11	734	1.59	1.45
vons5	147	0	839	14.91	0.00
vons6	665	8	2,258	22.69	0.27
Totals	**4,070**	**888**	**47,144**	**7.81**	**1.70**

between MW and LW. By detection distance, I mean the distance between the camera system and the object being detected. Given MW's shallower FoV, camera motion can be detected as a difference in the number of pixels in foreground between the MW and LW. The findings from this experiment are not published here due to not having labeled frames in sequences as moving/stationary. However, qualitatively this was an effective technique. Another technique investigated was to remove bounding boxes in moving images that did not contain bright pixels. This was an effective approach in scenes where the sky or clouds were the background with an airplane engine as the object or target being detected. Again given the lack of camera motion labels on the sequence, quantitative results were not calculated. It is likely that merely thresholding the video would have been more effective than using motion detection in these

Table 4.2: Comparison of DB IoU to MW IoU

Sequence	DB > MW	DB < MW	DB = MW	%-imp.	%-det.
brwncamp1	7,077	144	1,229	83.75	1.70
brwncamp2	4,600	195	760	82.81	3.51
brwncamp3	3,810	396	332	83.90	8.73
brwncamp5	5,268	902	637	77.39	13.25
brwncamp6	3,859	1,433	356	68.33	25.37
brwncamp7	2,984	625	352	75.33	15.78
brwncamp8	7,552	665	917	82.68	7.28
SBAP12	109	0	1	99.09	0.00
SBAP13	127	0	0	100.00	0.00
vons1	598	27	3	95.22	4.30
vons2	135	0	0	100.00	0.00
vons3	197	0	0	100.00	0.00
vons4	334	3	18	94.08	0.85
vons5	825	0	3	99.64	0.00
vons6	2,744	5	42	98.32	0.18
Totals	**40,219**	**4,395**	**4,650**	**81.64**	**8.92**

circumstances.

Tables 4.3 and 4.4 show the detection capabilities if IoU is set to a somewhat arbitrary value of 0.7330. Table 4.8 and Table 4.10 show the detection capabilities of YOLOv4 and YOLOv7 using a similar IoU threshold. In Tables 4.3 the per object detection information is given. There were very few cases were MWIR alone outperformed LWIR. Brwncamp4 object 8 had 4 MWIR detections to 2 LWIR and brwncamp9 object 5 had 21 MWIR ViBe detections to 5 LWIR ViBe detections. Brwncamp4 and brwncamp 9 are sequences where the camera was directed approximately perpindicular to the roadway, objects in these sequences are only available for a small number of frames. There is significant ego-motion in these sequences and the LWIR channel has significant false positive foreground detections. Since MWIR does not see distant objects

as well as LWIR with this camera setup the near foreground MWIR has fewer foreground false positives. Combining foreground false positives true positives cause the bounding box for brwncamp4 object 8 to be too large to be counted as a TP. The information provided in Table 4.3 is aggregated by object class and provided in Table 4.4. The only class where MWIR exceeded LWIR for ViBe detection is 'Semi and Trailer.' This class is well represented in brwncamp4 and brwncamp9 and MWIR outperformed LWIR for the same reasons as discussed for the objects in those sequences. There were 10 objects in Table 4.8 where the %-Improved was negative. In brwncamp3 there was a significant duration where object 3 and object 4 merge. They are traveling the same direction on a four lane road and object 4 overtakes object 3. Object 4 continues to occlude object 3 until they both are too small to continue labeling. Brwncamp4 object 7 is a pickup with a large box at an odd angle on the side. In the MWIR Brwncamp4 object 7 detects as two separate objects pickup and box. In the LWIR false positive common to brwncamp4 because of its ego-motion and background hills combine with MWIR foreground detections to make the bounding-box too large to be considered TP with our threshold. Brwncamp5 object 11 appears to be an occlusion event where two object and hence their bounding boxes merge. Brwncamp6 starts with several distant vehicles traveling together, the bounding boxes for these vehicles merge. Brwncamp7 objects 11 and 12 have several merging events with various other objects. In this sequence many vehicles are traveling in opposite directions with detections merging and then unmerging over the duration of the sequence. Brwncamp8 has similar levels of traffic conditions as brwncamp7 and objects 11, 12, and 13 have lower DBIR detections for the same reasons as objects 11 and 12 in brwncamp7. On this

data set there appear to be two main causes for decreased DBIR detection in ViBe. The first is occlusion detection merging and the second is ego-motion of the camera.

Directly comparing ViBe and YOLO is challenging due to the limited number of sequences with stationary FoV. However, Table 4.3 and Table 4.4 are calculated over the entire data set. One thing that can be inferred from these tables is that DBIR significantly improves object detection via motion by improving the connectivity of the object substantially. It does this across the entire data set but does better when the camera is stationary for the entire sequence.

Table 4.3: ViBE Motion Detection Results Per Object (IoU > 0.7330)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp1	1	SUV	35	26	35	0.00
brwncamp1	4	Car	168	47	213	21.13
brwncamp1	6	Car	13	1	13	0.00
brwncamp1	7	Car	320	35	322	0.62
brwncamp1	8	Car	186	11	186	0.00
brwncamp1	10	Car	158	17	186	15.05
brwncamp2	5	Pickup	145	14	148	2.03
brwncamp2	9	Car	73	7	82	10.98
brwncamp2	10	Car	136	4	177	23.16
brwncamp2	11	SUV	134	1	134	0.00
brwncamp3	2	Pickup	165	24	165	0.00
brwncamp3	3	Van	72	66	69	-4.35
brwncamp3	4	Car	35	18	35	0.00
brwncamp3	5	Pickup & Trailer	24	14	24	0.00
brwncamp3	7	SUV & Trailer	28	19	29	3.45
brwncamp3	8	SUV	73	17	73	0.00
brwncamp3	9	Car	111	34	112	0.89
brwncamp4	2	Semi & Trailer	5	5	5	0.00
brwncamp4	3	Semi & Trailer	35	28	36	2.78
brwncamp4	7	Pickup	10	3	9	-11.11
brwncamp4	8	Car	2	4	4	50.00
brwncamp5	1	Pickup	130	66	131	0.76
brwncamp5	2	Car	136	27	139	2.16
brwncamp5	3	Car	127	28	130	2.31
brwncamp5	6	Pickup	130	42	131	0.76
brwncamp5	7	Box Truck	130	108	130	0.00
brwncamp5	8	SUV & Trailer	113	1	113	0.00
brwncamp5	9	SUV & Trailer	123	5	123	0.00
brwncamp5	10	Semi	149	53	149	0.00
brwncamp5	11	Pickup	43	2	42	-2.38
brwncamp5	12	Pickup	50	2	51	1.96
brwncamp5	14	Car	38	4	46	17.39
brwncamp5	15	Pickup	51	20	55	7.27
brwncamp5	17	Van	12	30	13	7.69

Table 4.3: ViBE Motion Detection Results Per Object (IoU > 0.7330) (continued)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp6	1	Car	11	8	11	0.00
brwncamp6	3	Car	12	34	8	-50.00
brwncamp6	5	Indeterminate	7	1	8	12.50
brwncamp6	8	SUV	11	47	12	8.33
brwncamp6	9	Pickup	166	3	165	-0.61
brwncamp6	10	SUV	389	246	389	0.00
brwncamp6	11	Pickup	243	181	244	0.41
brwncamp6	12	SUV	215	114	224	4.02
brwncamp6	13	SUV	149	114	149	0.00
brwncamp6	15	Van	65	37	65	0.00
brwncamp7	1	SUV	23	1	26	11.54
brwncamp7	2	SUV	19	13	21	9.52
brwncamp7	3	Pickup	66	14	66	0.00
brwncamp7	5	Car	65	3	78	16.67
brwncamp7	6	Car	63	27	78	19.23
brwncamp7	7	Pickup	69	4	73	5.48
brwncamp7	10	Pickup	105	35	127	17.32
brwncamp7	11	Car	76	1	73	-4.11
brwncamp7	12	Pickup	158	67	150	-5.33
brwncamp7	13	Car	71	48	71	0.00
brwncamp7	14	Pickup	126	35	155	18.71
brwncamp7	15	Pickup	70	32	70	0.00
brwncamp8	4	Pickup	137	7	137	0.00
brwncamp8	5	Indeterminate	5	2	5	0.00
brwncamp8	6	SUV	172	15	172	0.00
brwncamp8	9	Car	19	12	23	17.39
brwncamp8	11	Pickup	108	43	107	-0.93
brwncamp8	12	Pickup	30	6	28	-7.14
brwncamp8	13	Pickup	29	16	25	-16.00
brwncamp8	14	Car	77	12	77	0.00
brwncamp8	15	Pickup	86	18	86	0.00
brwncamp8	20	Pickup	45	2	45	0.00

Table 4.3: ViBE Motion Detection Results Per Object (IoU > 0.7330) (continued)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp9	1	Pickup	10	1	11	9.09
brwncamp9	3	SUV	6	4	7	14.29
brwncamp9	5	Semi & Trailer	5	21	32	84.38
brwncamp9	7	Car	2	1	3	33.33
brwncamp9	9	Car	5	3	5	0.00
brwncamp9	11	SUV	5	3	5	0.00
brwncamp9	12	Pickup	7	5	8	12.50
brwncamp9	13	Pickup	5	3	5	0.00
brwncamp9	14	Car	16	12	16	0.00
brwncamp9	15	Car	6	5	8	25.00
brwncamp9	16	Car	3	1	4	25.00
SBAP12	2	Car	50	5	52	3.85
SBAP17	3	SUV	89	25	90	1.11
SBAP17	4	SUV	43	20	43	0.00
SBAP26	1	Airplane	21	4	40	47.50
vons1	3	Person	14	6	14	0.00
vons1	7	Pickup	2	2	2	0.00
Sum	**All**		**6,336**	**2,102**	**6,623**	**4.52%**

Table 4.4: ViBe Motion Detection by Object Class (IoU > 0.7330)

Class	LW Num	MW Num	DB Num	%-Imp
Airplane	21	4	40	47.50
Box Truck	130	108	130	0.00
Car	1,979	409	2,152	8.04
Indeterminate	12	3	13	7.69
Person	14	6	14	0.00
Pickup	2,186	647	2,236	2.24
Pickup & Trailer	24	14	24	0.00
Semi	149	53	149	0.00
Semi & Trailer	45	54	73	38.36
SUV	1,363	646	1,380	1.23
SUV & Trailer	264	25	265	0.38
Van	149	133	147	-1.36

4.1.4 YOLOv4 & v7 Experiments

Given the recency and fast moving nature of the changes to the YOLO architecture, YOLOv4 and YOLOv7 were selected as test CNNs for experimentation. YOLOv4 and v7 are readily available pretrained on MS COCO and PASCAL VOC datasets. Models pretrained on MS COCO were selected and obtained. Inference was then conducted on the DBIR dataset presented in this work.e

To establish the benefit of preprocessing, YOLOv4 was run against both raw and preprocessed images. The results are tabulated below. Given the level of benefit of preprocessing, YOLOv7 was only evaluated against the preprocessed data set.

Preprocessing Images for CNN Inference

Initial testing involved passing images that were unaltered, other than the scaling that occurs when the image was transformed into a native image format.

It was observed that YOLOv4 performed poorly when there was a highly emissive object in the scene. In Fig. 4.7, the unprocessed histogram for the pixel values of Brwncamp1 in LW is displayed. Given the sensitivity of the FPA to temperature parts of the detector would occasionally "freeze out." This was caused by imprecise control of the pressures required to cool the FPA to approximately 77 K. If the FPA got too cool, some pixels would fail to respond to input and stop reporting valid data. Figs. 4.9 and 4.10 show the histogram with the stuck and frozen pixels removed. Figs. 4.11 and 4.12 show the LWIR and MWIR histograms overlaid with a Gaussian Mixture Model (GMM) fitted to the histogram. In black the sample surprise was plotted.

For the purposes of running YOLOv4 and YOLOv7, it proved not particularly necessary to remove the frozen or stuck pixels from the data set. However, the scale of the spike at pixel intensity of 3,194 in Fig. 4.7, interferes with some of the statistical analyses that follow. The value of 3,194 is not universal on all sequences, and the value of the bin with the maximum count from the histogram was removed. The frozen-pixel spike is removed in Fig. 4.9. It is important to take note of the long right (hot) tails of the histograms in Figs. 4.7, 4.9, 4.8,and 4.10. For the MW about 8.8% of the pixels were damaged prior to acquisition of the FPA. The damaged FPA was accepted for cost saving purposes. The damaged pixel always return a pixel intensity of 3,084. The spike at 3,084 can be seen in Fig. 4.8 and the histogram after its removal in Fig. 4.10.

Given that Off-the-Shelf (OTS) versions of CNNs, including YOLOv4, require standardized inputs, these images will have to be scaled to the required input values for the NN architecture. Full-scale stretching of these images

as-is creates very bright spots for the intensity values on the right, with the remainder of the image being very dim. This packs most of the information content of the image, from the perspective of an OTS CNN, into very few values. This results in a significant information loss and poorer performance.

To improve the performance of CNN on this data set a somewhat simple yet effective strategy was devised to improve histogram equalization. Initial investigations into using log-scaling were investigated but the overall brightness of the images still changed significantly frame-to-frame. The impact of this on the CNN are unclear without further research. However, a qualitative decision was made to clip pixel values from the low and the high sides of the distribution such that the minimum threshold was the smallest intensity value with at least 1,000 occurrences, and the maximum was the largest value exceeding a bin count of 1,000. Distant objects in the MWIR, the spectrum which was most required scaling or thresholding, appeared dimmer and harder to distinguish using *log*-thresholded values relative to bright foreground objects. Fig. 4.6 shows a side-by-side comparison of my heuristic approach and *log*-scaling. After some initial testing I decided that my heuristic approach met my needs and did not persue perfecting a *log*-based approach. Values below the minimum value were set to the minimum value, and similarly, values above the maximum were set to the maximum. In Fig. 4.13 this results in a large number of overall pixels being set to the maximum threshold. This resulted in stable image brightness and retained detail in the images.

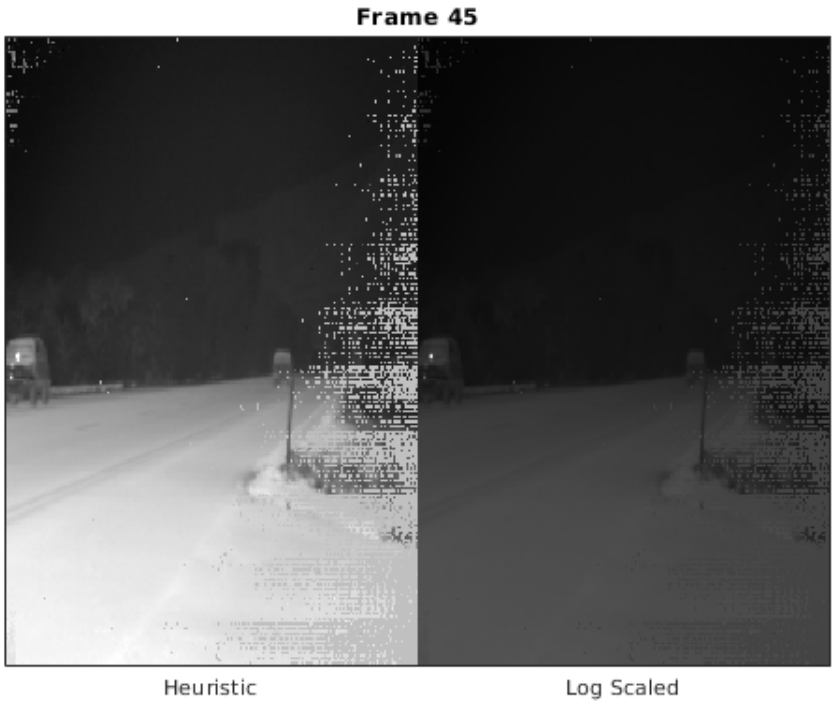


Figure 4.6: Side-by-side comparison of log-scaling and heuristic thresholding.

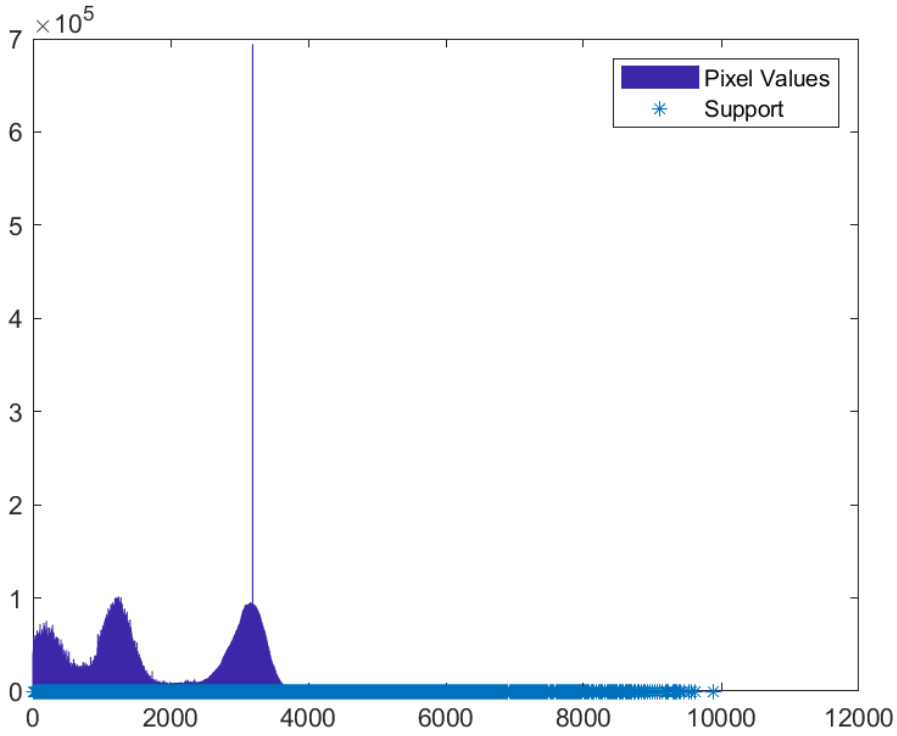


Figure 4.7: Brwncamp1 LW pixel value histogram before thresholding process. Dark blue values represent the pixel value counts for each value. Light blue x's display all non-zero values. Frozen pixel spike at pixel intensity 3,194.

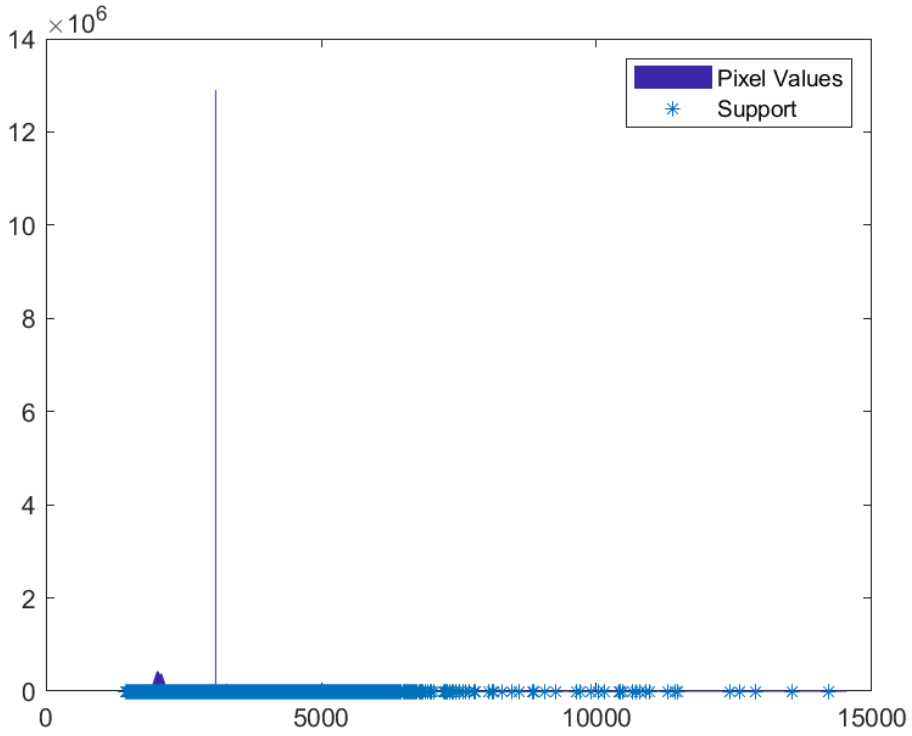


Figure 4.8: Brwncamp1 MW pixel value histogram before thresholding process. Dark blue values represent the pixel value counts for each value. Light blue x's display all non-zero values.

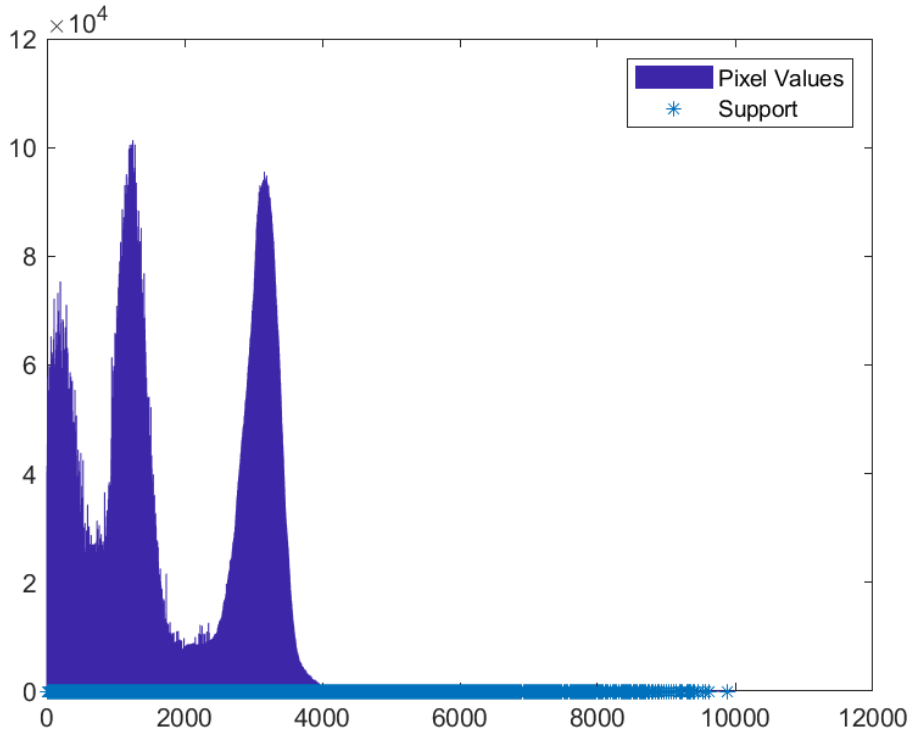


Figure 4.9: Brwncamp1 LW pixel value histogram with "frozen pixel" spike removed. Dark blue values represent the pixel value counts for each value. Light blue x's display all non-zero values.

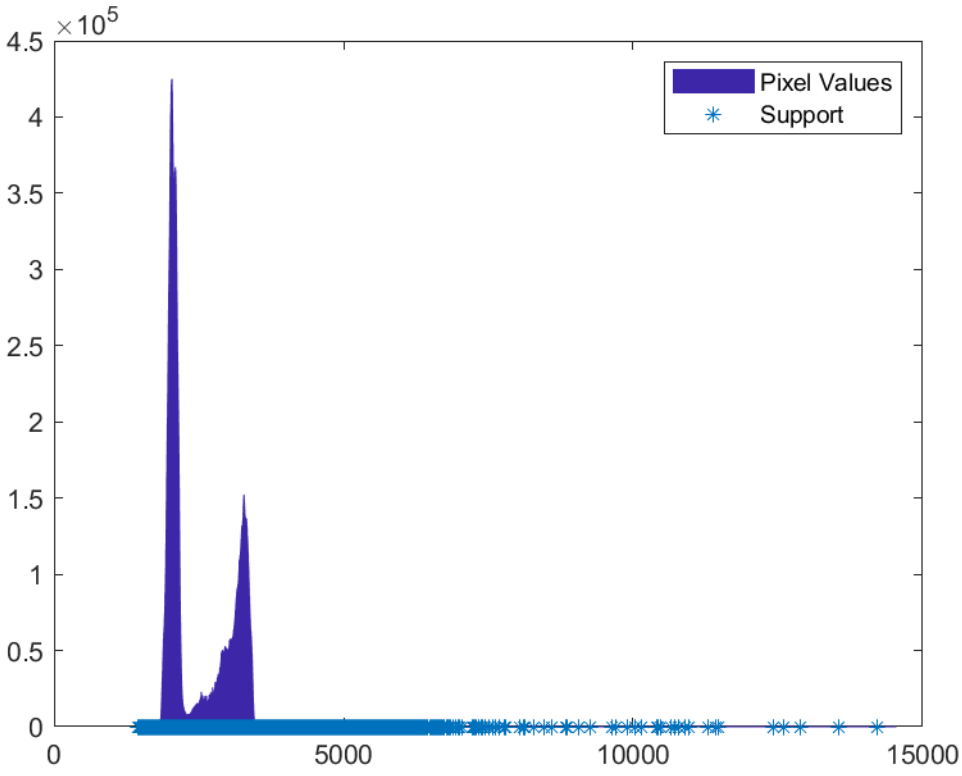


Figure 4.10: Brwncamp1 MW pixel value histogram with "damaged pixel" spike removed. Dark blue values represent the pixel value counts for each value. Light blue x's display all non-zero values.

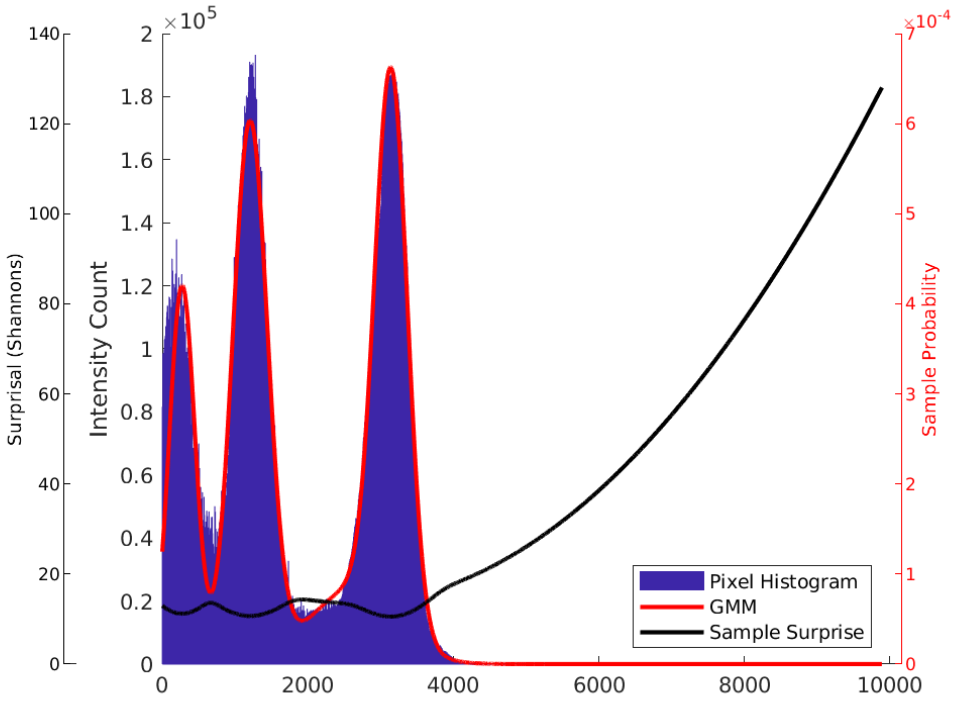


Figure 4.11: Brwncamp1 LW histogram in blue, Sample Probabilty from Gaussian Mixture Model in red, and the surprisal value in shannons in black.

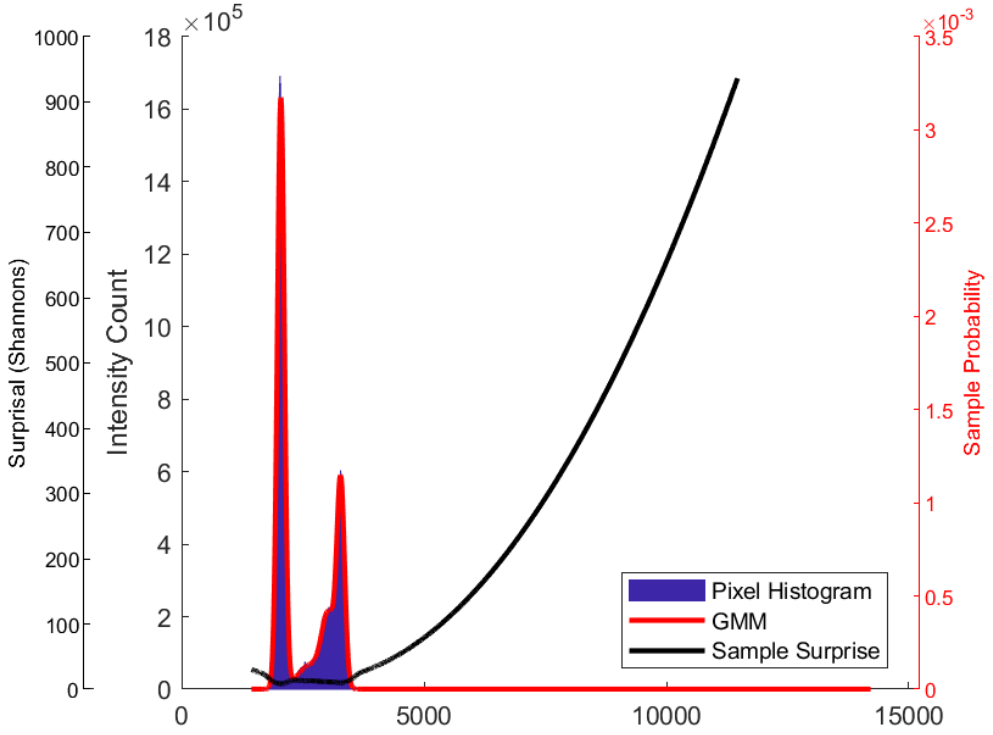


Figure 4.12: Brwncamp1 MW histogram in blue, Sample Probabilty from Gaussian Mixture Model in red, and the surprisal value in shannons in black.

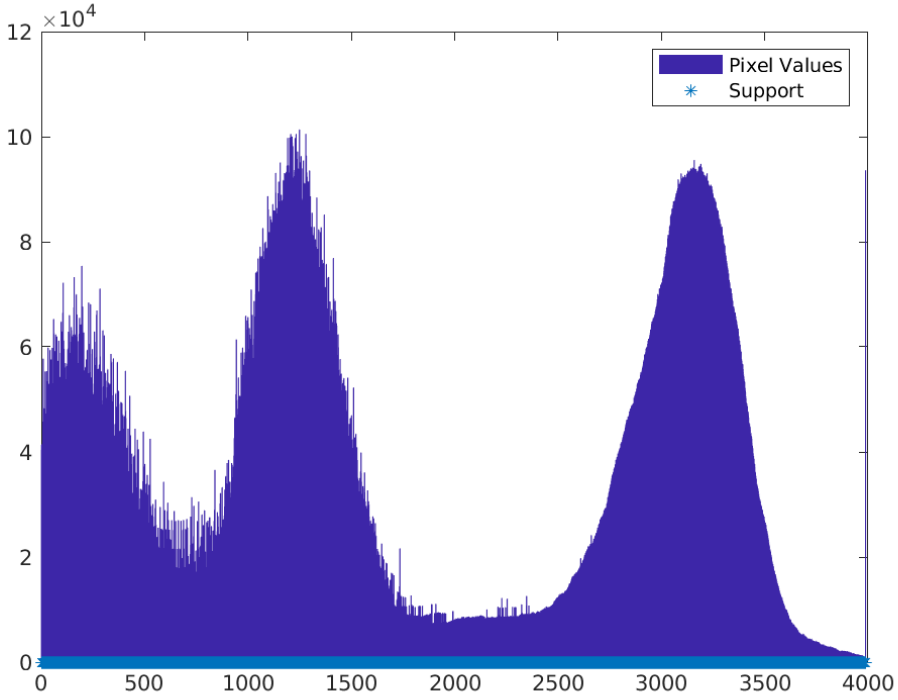


Figure 4.13: Brwncamp1 LW pixel value histogram after thresholding.

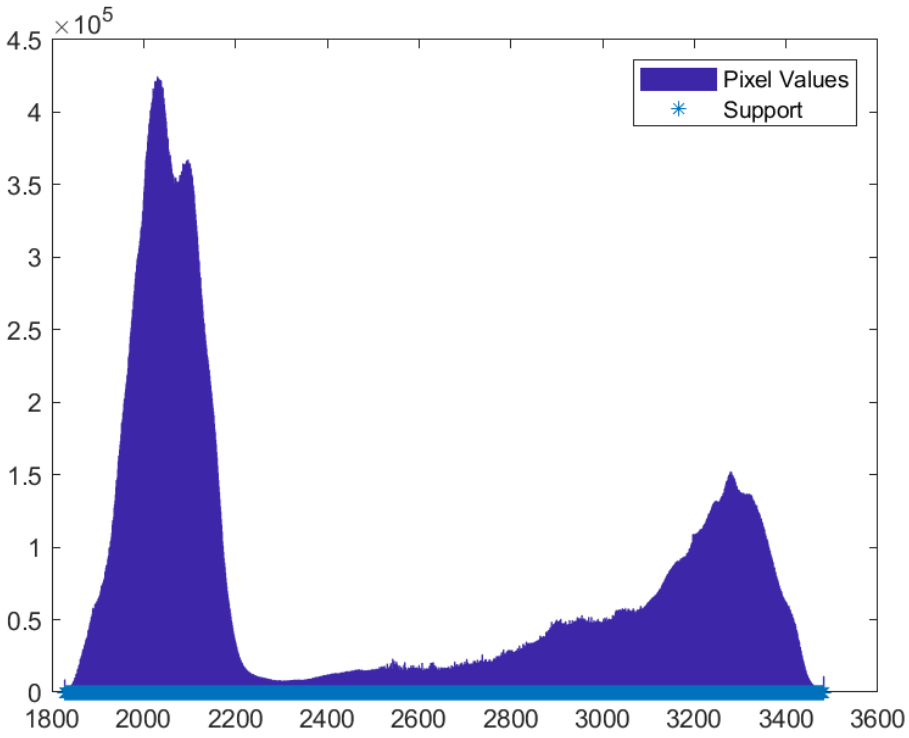


Figure 4.14: Brwncamp1 MW pixel value histogram after thresholding.

Fig. 4.13 shows the histogram of Brwncamp1 LW after thresholding. The LW data is not significantly offset from zero, as can be seen in Figs. 4.7 and 4.11. Although the lower threshold is relatively unimportant in LW, it is much more important in the MW, as seen in Figs. 4.12 and 4.14. Some non-zero minimal thresholds for LW were observed though as can be seen in the "LW Min" column of Table 4.5. This is specifically due to longer tails in MW and the fact that there was an offset in the Data Acquisition Device (DAQ). All sequences in this work were then analyzed for appropriate thresholding values listed in Table 4.5. This *ad hoc* method has proven useful for my research purposes but is not suggested to be "optimal." Further research into optimizing these thresholds with respect to a target CNN could be achieved by searching the threshold space using the detection performance of the CNN as a loss function.

In an attempt to derive an Information Theory based thresholding algorithm I took the histograms of the sequences as seen in Figs. 4.7 and 4.8 and fit GMMs to the data. This process attempts to fit multiple Gaussians to better model a multi-modal distribution. My thinking was that I would take the GMMs in Fig. 4.13 and 4.14 and look at the constituent Gaussian with the highest and lowest modes. The maximal threshold could be set to some fixed number of standard deviations above the maximal Gaussian and similarly a fixed number of standard deviations below the minimal Gaussian, although initial investigation into this process proved resistant to solution, with situations of too much information or too little information occurring seemingly randomly.

The three modes represented by the peaks of the un-normalized distribu-

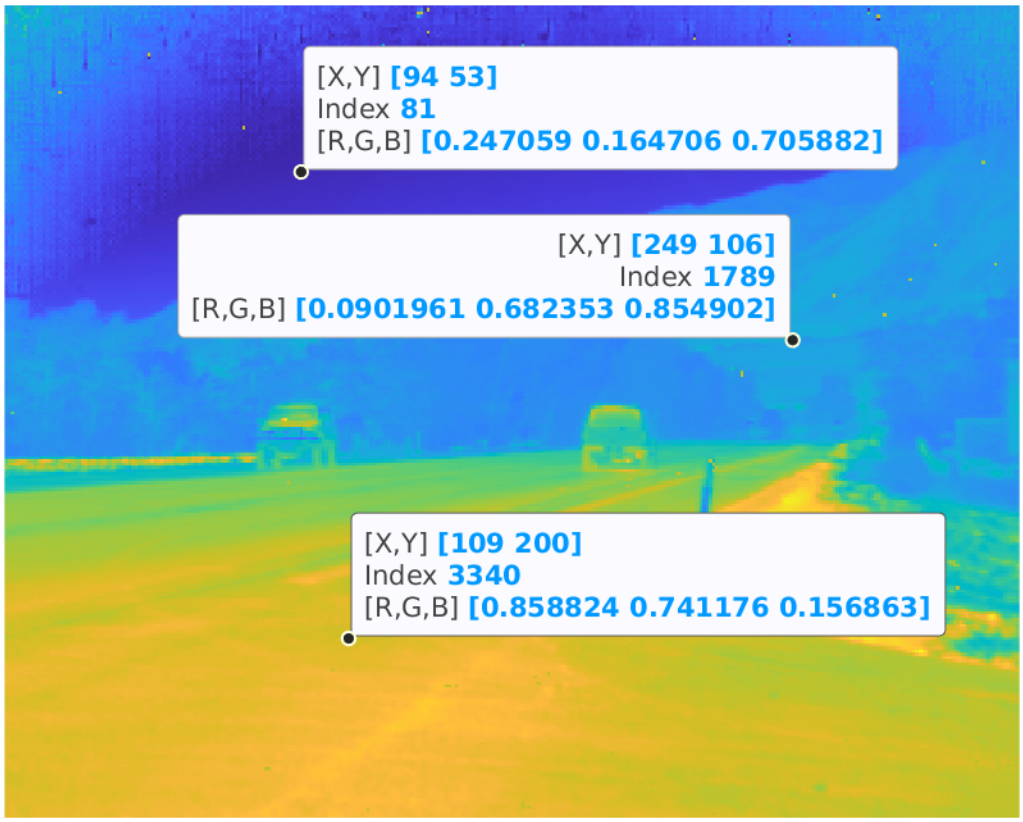


Figure 4.15: Brwncamp1 LW image with labels at portions of the scene corresponding to modes in the pixel intensity histogram.

tions in 4.13 correspond to sky, background terrain, and foreground pavement from left to right. As an illustration, Fig. 4.15 shows the image background with labels added to the foreground. The "index" in the labels represents the pixel intensity at that location in the image. These distributions are contingent on the scene. Many of the brwncamp sequences have a three mode distribution similar to Fig. 4.13 and 4.14. However, changes of scene significantly change the pixel intensity distribution. To model these distributions each was fitted with a Gaussian Mixture Model (GMM), similar to [123] among others, to the pixel intensity distributions.

To understand how much information content each pixel contains we can use surprisal analysis. If $pdf(\cdot)$ represents the probability distribution function represented by the GMM then surprisal is calculated as:

$$S(\cdot) = -\log_b(pdf(\cdot)). \tag{4.1}$$

The base of the logarithm in Eq. 4.1 determines the units. When the base b is 2, the units are "Shannons" or "bits". If the base is Euler's number e, then the unit is "nats," and similarly, "hartley" for base 10 [63]. Surprisal is devised to quantify the amount of information in a given value relative to the probability distribution of the analyzed variable. In terms of pixel intensity, thresholding the distributions in Fig. 4.11 and 4.12 removes the pixels intensities with the most information content, *i.e.*, the most "surprising" pixels. The CNN-based approaches, YOLO *et. al* depend on the information between pixels rather than the content of single pixels. It can not be concluded, however, that training a CNN on the full range of the new DBIR dataset with the full range of values wouldn't improve detection capabilities of that CNN.

Table 4.5: Thresholding Values

Sequence	LW Min	LW Max	MW Min	MW Max
brwncamp1	0	3,988	1,828	3,482
brwncamp2	0	3,568	1,731	3,364
brwncamp3	0	4,414	2,271	3,721
brwncamp4	0	3,290	2,166	3,646
brwncamp5	0	4,723	2,249	4,139
brwncamp6	0	4,189	2,286	4,198
brwncamp7	0	4,081	2,274	4,158
brwncamp8	0	4,393	2,192	4,138
brwncamp9	0	2,364	2,056	3,134
SBAP1	1,348	3,148	8,204	9,274
SBAP2	1,200	2,538	8,058	8,770
SBAP3	1,522	2,766	8,111	8,906
SBAP5	4,922	8,269	8,626	1,0163
SBAP6	5,064	8,316	8,670	1,0232
SBAP7	4,038	7,991	8,468	1,0152
SBAP8	2,616	7,991	7,348	1,0156
SBAP9	4,200	6,330	8,214	9,042
SBAP10	3,214	6,042	7,940	9,045
SBAP11	2,350	6,042	7,946	9,048
SBAP12	2,948	5,322	7,420	8,161
SBAP13	2,938	5,730	7,394	8,509
SBAP14	2,572	5,061	6,776	8,144
SBAP15	2,593	5,127	6,828	8,144
SBAP16	0	2,510	5,221	6,430
SBAP17	0	2,401	5,164	6,430
SBAP18	0	2,353	5,054	6,430
SBAP19	0	2,401	4,805	6,430
SBAP20	0	2,712	4,805	6,430

Table 4.5: Thresholding Values (Cont.)

Sequence	LW Min	LW Max	MW Min	MW Max
SBAP21	1,218	4,299	6,162	7,235
SBAP22	207	3,425	5,809	6,783
SBAP23	0	3,081	1,828	6,805
SBAP24	119	3,289	5,367	6,611
SBAP25	0	2,668	5,316	6,430
SBAP26	0	2,353	4,541	6,430
SBAP27	920	3,901	5,955	7,043
SBAP28	355	3,515	5,969	6,838
vons1	0	2,412	1,455	2,127
vons2	0	1,936	1,900	2,262
vons3	0	1,802	1,900	2,283
vons4	0	1,794	1,530	2,198
vons5	0	2,739	3,012	3,608
vons6	0	2,518	2,112	3,567

In Table 4.6, the effects of thresholding on pixel-intensity information content are presented. A graph of a sample surprise value function can be seen in Fig. 4.11 and Fig. 4.12. Retained information was calculated as

$$Retained = \sum_{\substack{k \in P \\ l<k<u}} S(k) \tag{4.2}$$

Here, P represents the pixel intensity values of each pixel in the video sequence, l is the lower threshold, u is the upper threshold, and S is defined in (4.1). To calculate the rejected information the following equation was used:

$$Rejected = \sum_{\substack{k \in P \\ k>u}} S(k) + \sum_{\substack{k \in P \\ k<l}} S(k) \tag{4.3}$$

. Then I calculated the "Rejected %" as:

$$Rejected\% = \frac{Rejected}{Retained + Rejected} \tag{4.4}$$

Table 4.6: Thresholding Effect on Pixel-wise Information Content

Sequence	LW Retained	LW Rejected	LW Rejected %	MW Retained	MW Rejected	MW Rejected %
brwncamp1	1.65E9	5.74E7	3.36	1.47E9	4.10E5	0.03
brwncamp2	1.05E9	9.05E7	7.95	9.72E8	9.73E5	0.10
brwncamp3	1.04E9	1.34E8	11.36	9.93E8	6.10E5	0.06
brwncamp4	1.03E9	1.13E8	9.89	9.66E8	2.49E6	0.26
brwncamp5	1.13E9	6.15E7	5.17	1.03E9	1.42E6	0.14
brwncamp6	1.12E9	2.31E7	2.03	1.03E9	1.06E6	0.10
brwncamp7	1.12E9	2.53E7	2.21	1.03E9	1.49E6	0.14
brwncamp8	1.11E9	6.88E7	5.82	1.04E9	9.17E5	0.09
brwncamp9	8.49E8	2.40E8	22.05	9.18E8	3.87E6	0.42
SBAP1	1.36E8	2.37E6	1.71	1.26E8	5.86E5	0.46
SBAP2	7.59E8	2.62E6	0.34	6.75E8	1.56E6	0.23
SBAP3	3.86E8	4.19E6	1.07	3.47E8	9.48E5	0.27
SBAP5	5.98E8	1.82E6	0.30	5.35E8	9.75E5	0.18
SBAP6	5.91E8	2.71E6	0.46	5.18E8	7.22E5	0.14
SBAP7	5.89E8	7.45E6	1.25	5.12E8	8.20E5	0.16
SBAP8	6.14E8	1.13E7	1.80	5.51E8	1.25E6	0.23
SBAP9	5.62E8	3.08E6	0.54	4.96E8	3.37E5	0.07
SBAP10	5.69E8	8.45E6	1.46	4.92E8	5.64E5	0.11
SBAP11	5.52E8	6.14E6	1.10	4.79E8	4.90E5	0.10
SBAP12	2.79E8	5.91E6	2.07	2.44E8	1.53E6	0.62
SBAP13	2.83E8	3.17E6	1.11	2.46E8	1.21E6	0.49
SBAP14	2.82E8	3.33E6	1.17	2.32E8	3.15E7	11.97
SBAP15	2.88E8	2.23E6	0.77	2.34E8	3.19E7	11.99
SBAP16	4.27E8	4.76E6	1.10	3.42E8	5.25E7	13.32
SBAP17	4.23E8	5.32E6	1.24	3.41E8	5.31E7	13.47
SBAP18	4.15E8	8.83E6	2.08	3.39E8	5.36E7	13.65
SBAP19	4.11E8	1.65E7	3.87	3.49E8	6.02E7	14.71
SBAP20	9.31E8	6.73E7	6.74	8.13E8	1.18E8	12.65

Table 4.6: Thresholding Effect on Pixel-wise Information Content

Sequence	LW Retained	LW Rejected	LW Rejected %	MW Retained	MW Rejected	MW Rejected %
SBAP21	9.84E8	2.78E6	0.28	8.18E8	2.21E6	0.27
SBAP22	5.89E8	2.58E6	0.44	4.84E8	1.54E6	0.32
SBAP23	5.65E8	9.50E6	1.65	1.58E9	1.50E6	0.09
SBAP24	5.92E8	2.22E6	0.37	5.23E8	1.04E6	0.20
SBAP25	5.58E8	2.46E7	4.23	4.34E8	6.87E7	13.66
SBAP26	5.31E8	5.86E7	9.93	4.89E8	7.06E7	12.62
SBAP27	1.16E9	2.92E6	0.25	9.77E8	1.23E6	0.13
SBAP28	1.14E9	2.64E6	0.23	9.17E8	1.55E6	0.17
vons1	2.08E9	1.70E7	0.81	1.60E9	2.23E6	0.14
vons2	1.00E9	1.54E7	1.51	8.02E8	3.63E5	0.05
vons3	9.23E8	1.27E8	12.11	8.04E8	2.07E6	0.26
vons4	9.36E8	1.00E8	9.68	7.63E8	1.46E8	16.02
vons5	1.49E9	5.99E6	0.40	1.36E9	2.41E6	0.18
vons6	1.49E9	6.47E5	0.04	1.27E9	3.47E5	0.03

.

Across all sequences, the total rejected percent in LW is 3.91%, with a standard deviation of 4.50%, and in MW, approximately 2.5%, with a standard deviation of 5.72%. The reader is reminded that the pixel intensity information is only part of the total information content of an image and that the spatial spatial relationships between the pixels also embeds information. As an example of exploiting this information, one of the recent leading contenders for modeling appearance was the Histogram-of-Oriented-Gradients (HoG) [38], which uses local spatial information in the form of gradients in the pixel intensity within a local window. Thresholding was not used in the ViBe experiments as it was beneficial to detection to allow foreground values to be extremely distant to background samples.

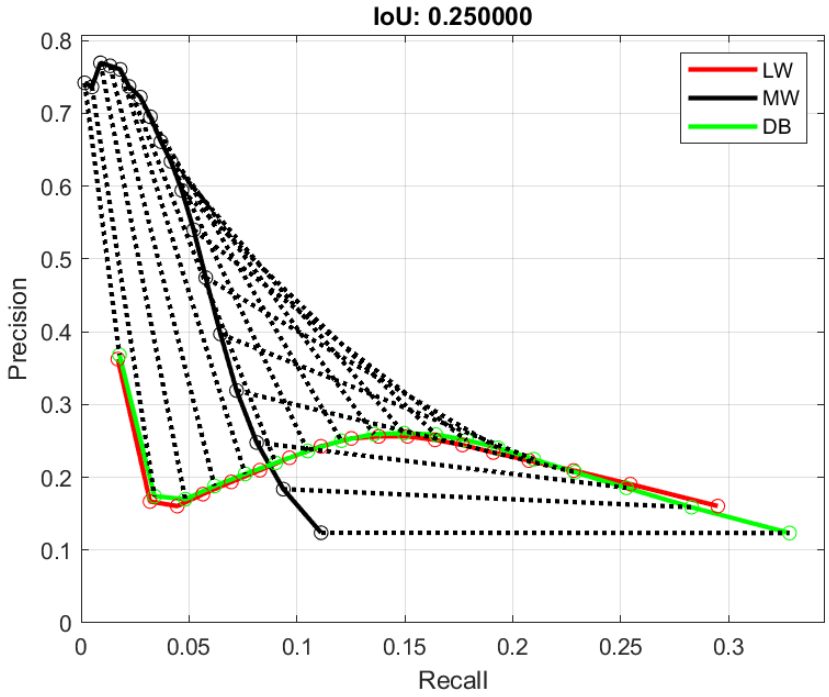


Figure 4.16: YOLOv4: Precision-Recall Curves for IoU 0.25. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

This thresholding resulted in a significant improvement in the performance of YOLOv4 on this data set, as shown in Tables 4.7 and 4.8. The total LW detections on unthreasholded images is totaled at the bottom of Table 4.7, and thresholded on the bottom line of Table 4.8. These tables shows that in LW 12,432 object instances were detected on thresholded images compared 12,102 in unthresholded. In LW, the total increase of detections between the unthresholded and thresholded is 330 detections for an improvement of 2.7% at IoU of 0.7330. Similarly these same tables show that in MW, the increase is 2,546 detections, corresponding to a 37.5% improvement at IoU of 0.7330. For

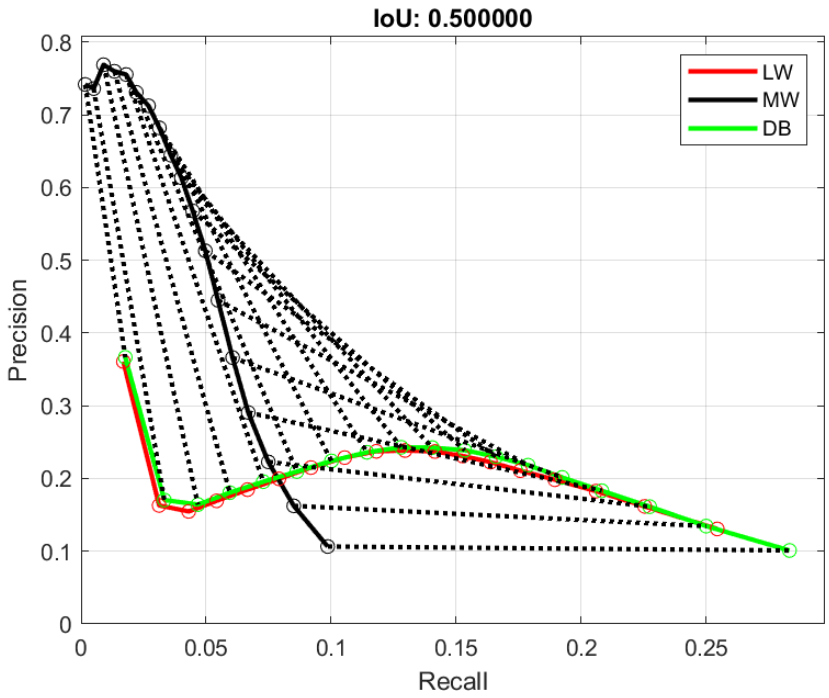


Figure 4.17: YOLOv4: Precision-Recall Curves for IoU 0.50. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

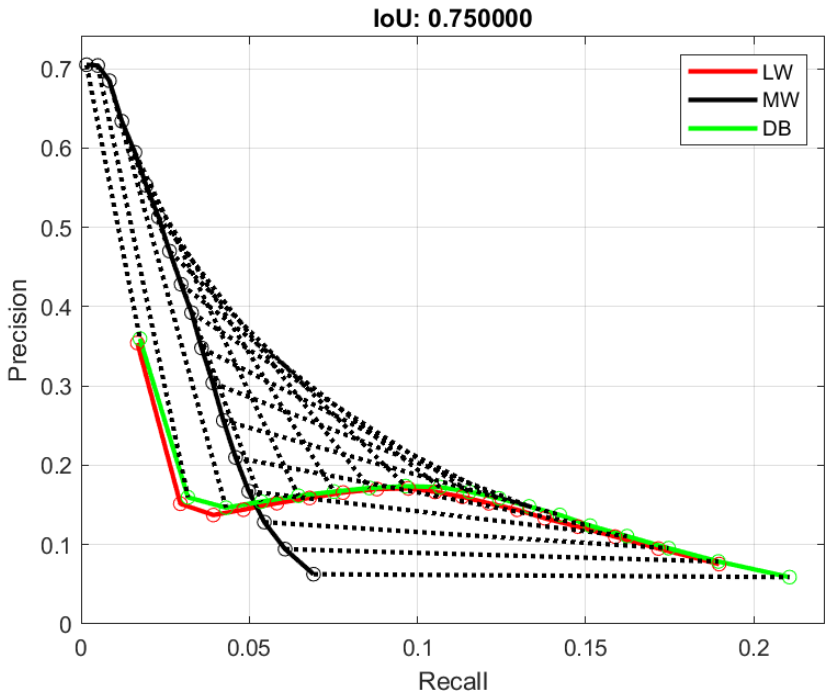


Figure 4.18: YOLOv4: Precision-Recall Curves for IoU 0.75. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

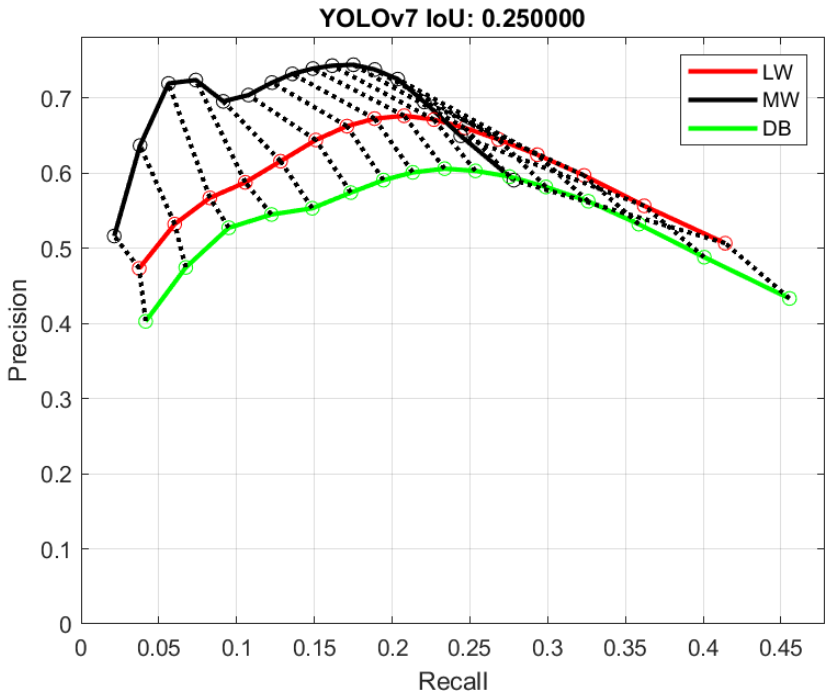


Figure 4.19: YOLOv7: Precision-Recall Curves for IoU 0.25. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

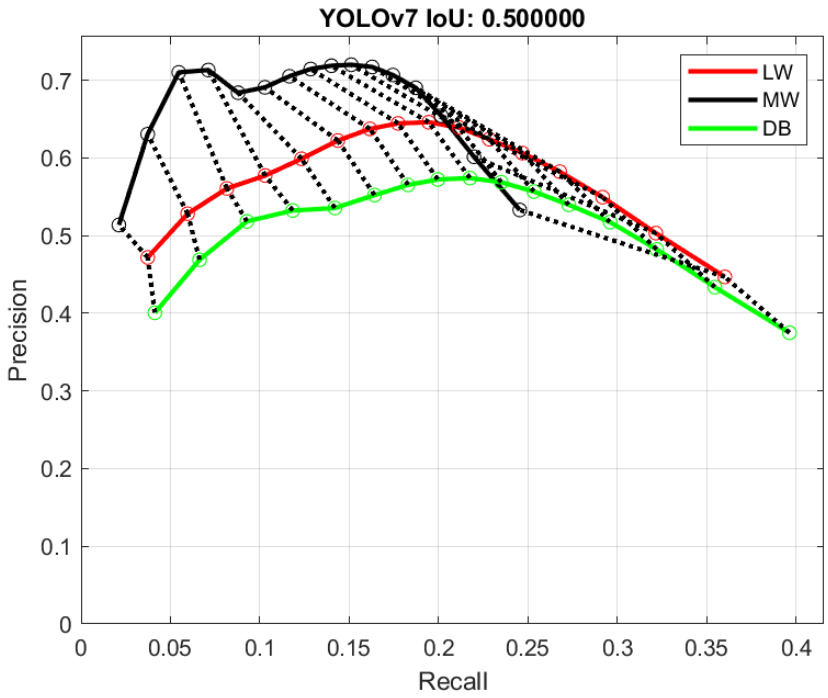


Figure 4.20: YOLOv7: Precision-Recall Curves for IoU 0.50. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

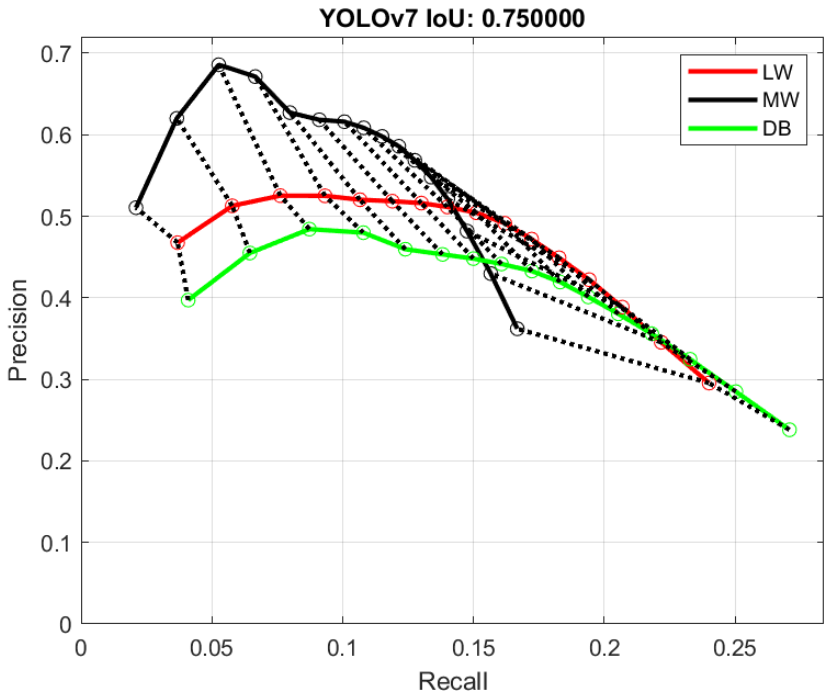


Figure 4.21: YOLOv7: Precision-Recall Curves for IoU 0.75. Dotted lines connect the MW to the DB for each classification probability 0.05-0.95. The green line is the DB curve and the red line is the LW. Note that the recall of the DB exceeds or is equal to the recall of either LW or MW alone and that the precision of the MW improves the precision of DB over LW.

DBIR the improvement from thresholding is 1,378 detections for approximately 10.6%. It can be concluded that in the case of this data set, preprocessing improved performance of YOLOv4.

Figures 4.22, 4.23, and 4.24 demonstrate the primary mode of improvement for YOLOv4 and YOLOv7 using two spectra. My observation was that these CNNs tend exhibit binary behavior in their detections, that is, the object is either detected or not detected at all. For some reason a CNN would detected well on one frame and then not detect the same object in subsequent frames. This was typically observed when the object was becoming closer to the camera and the size and appearance information of the object was increasing. This is likely due to the small and distant objects issues discussed in Chapter 2. I am not saying that the CNNs did not detect partial objects, because partial detection was observed. However, ViBe, in comparison is shown in some cases to have a decrease in performance when two spectra are used. This can be observed as negative %-Improved in Table 4.3 and Table 4.4. Since we are not unioning the bounding boxes from the CNNs in this experiment there is not a mechanism for one detection from the CNN to distort the size of another. In ViBe two moving objects occlude or nearly occlude the objects can be merged to form a larger detection bounding box now encompassing both objects.

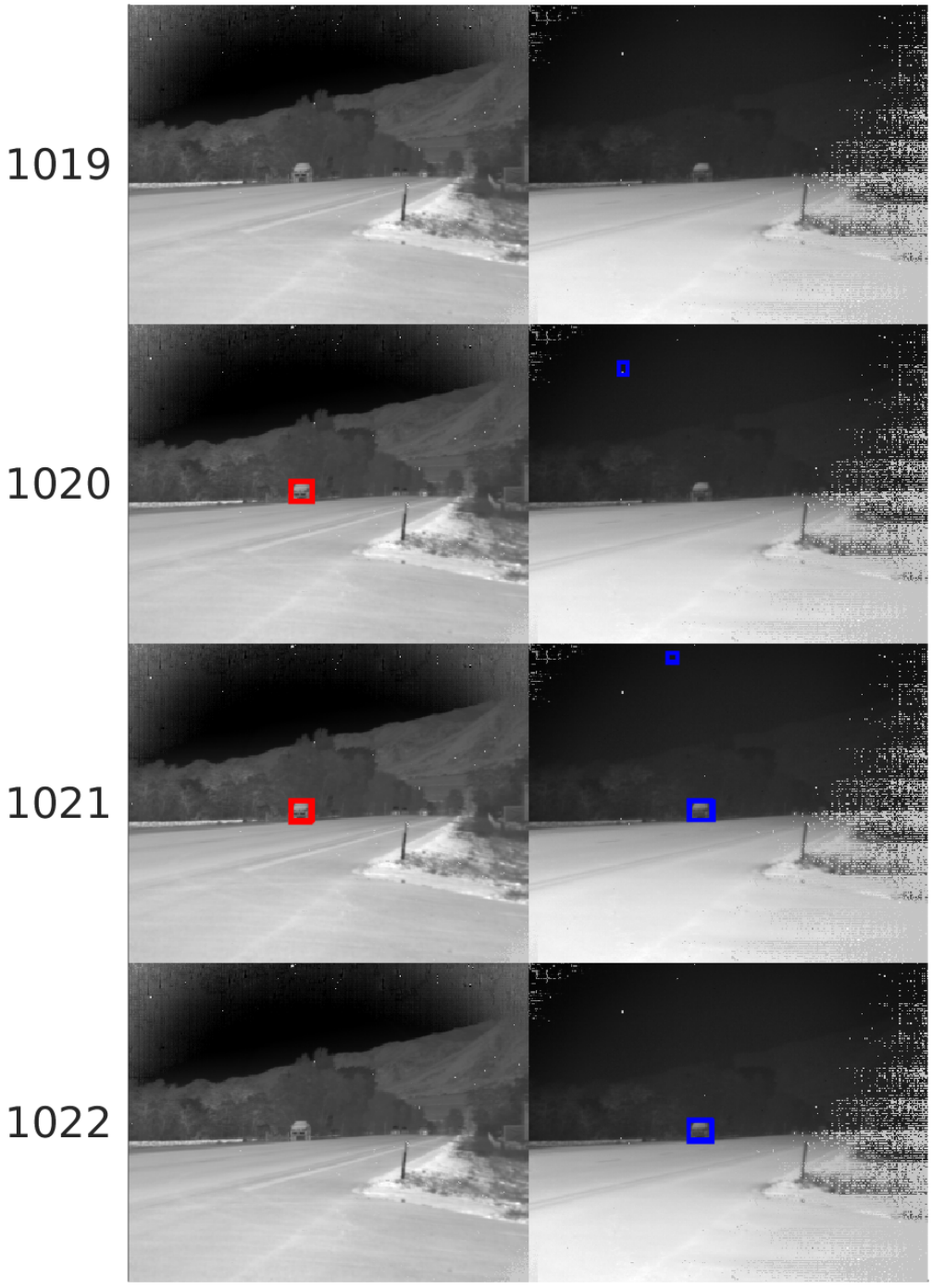


Figure 4.22: Frames 1,019 through 1,022 of sequence brwncamp1. Left column Long-wave with red detection bounding boxes. Right column is mid-wave with blue detection bounding boxes

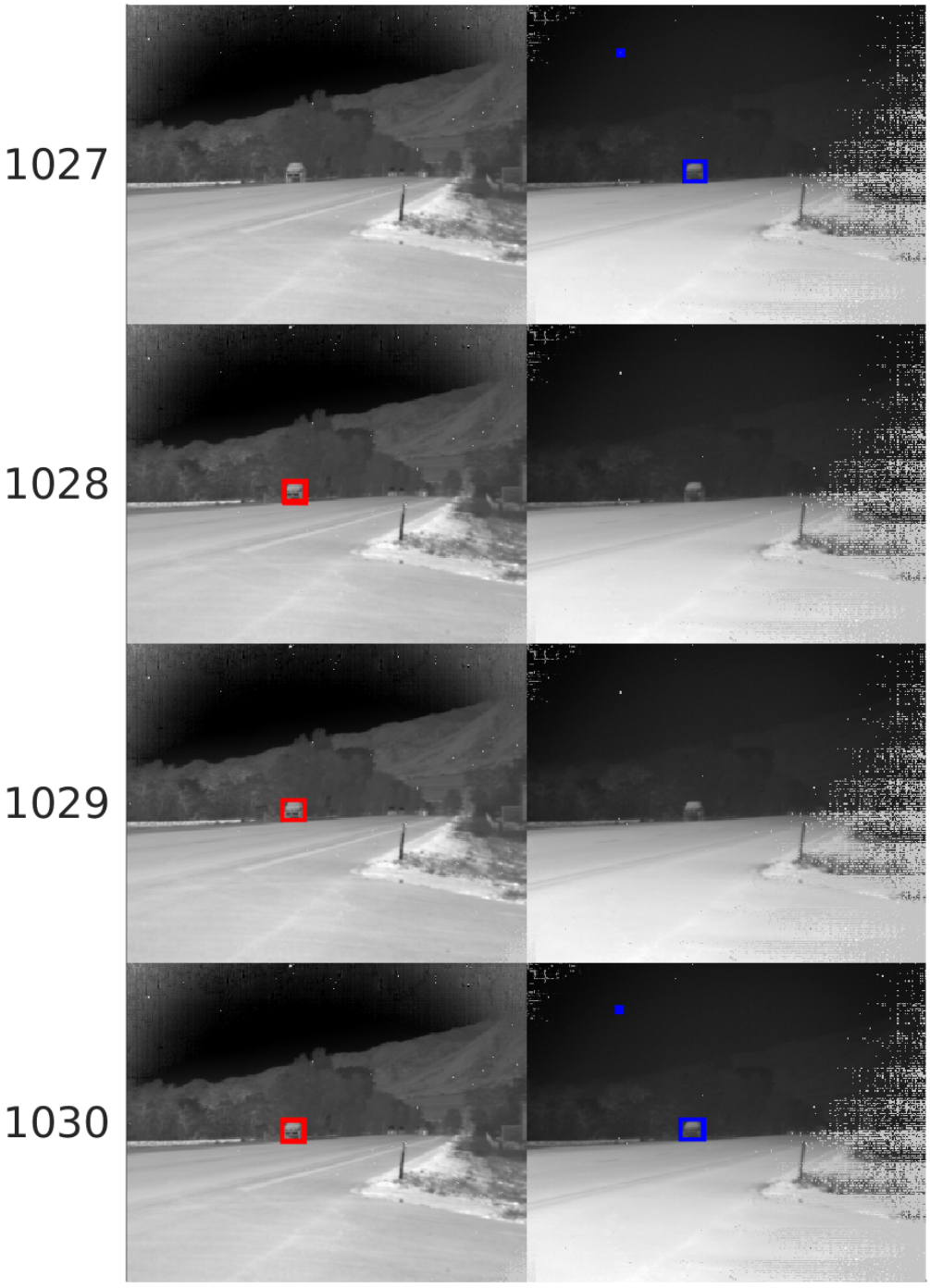


Figure 4.23: Frames 1,027 through 1,030 of sequence brwncamp1. Left column Long-wave with red detection bounding boxes. Right column is mid-wave with blue detection bounding boxes

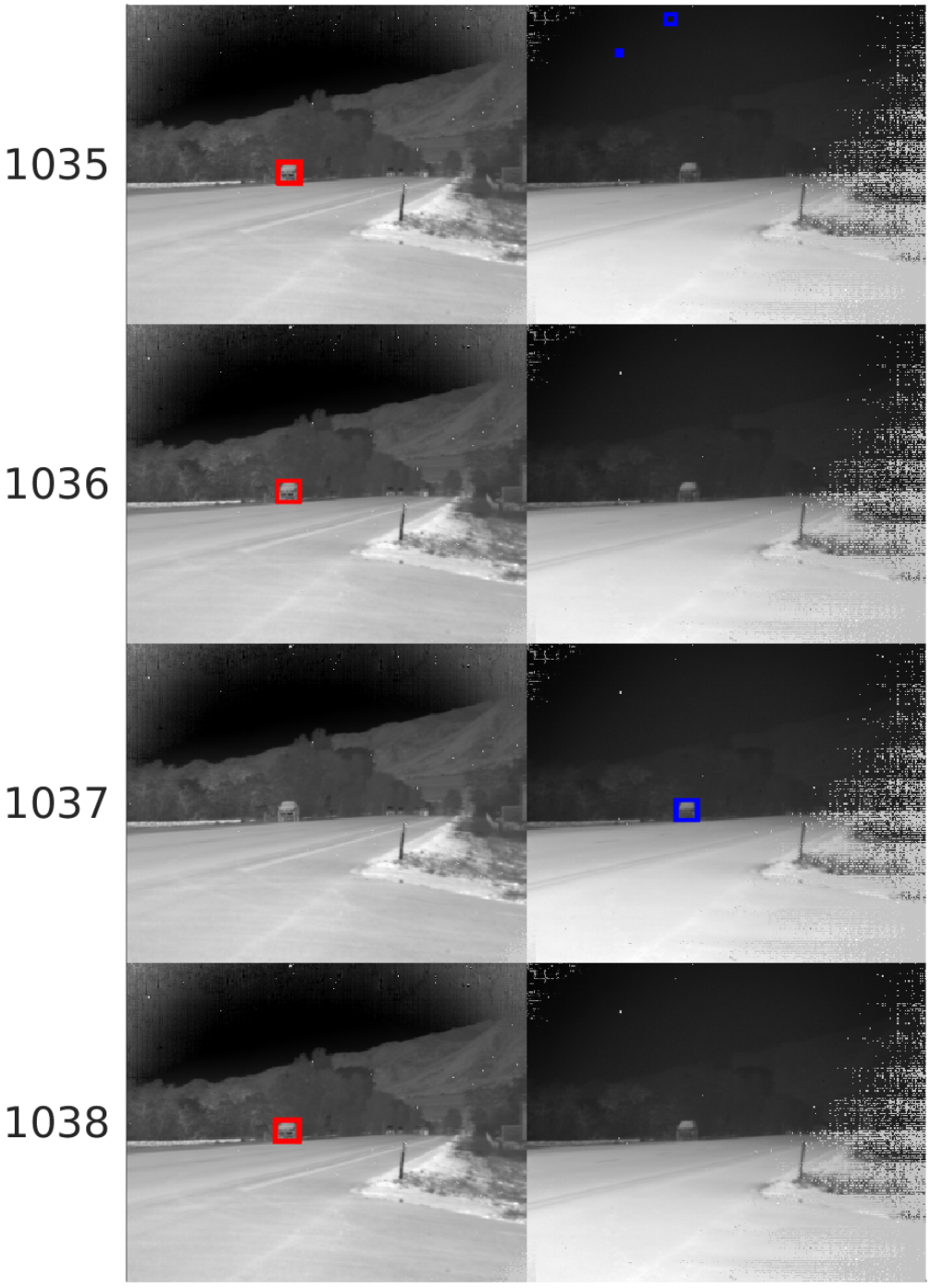


Figure 4.24: Frames 1,035 through 1,038 of sequence brwncamp1. Left column Long-wave with red detection bounding boxes. Right column is mid-wave with blue detection bounding boxes

In Table 4.7 the TP counts are aggregated by object. The number of LW, MW and DBIR detections per sequence are listed for each object in each sequence. What can be observed is that largest percent improvements are for objects with low detection counts. For example SBAP12 Object 2 was detected twice in total. Some objects had no detections that were not detected in both LW and MW, such as vons5 Object 2 and vons5 Object 3. On the other hand vons3 Object 2 held no detections common between MWIR and LWIR. Brwncamp 8 saw a significant increase in detections where 75% of MWIR detections were distinct from LWIR detections. There were no objects that were detected by LWIR that were not also detected by MWIR at some point in the sequence. Comparing the per object results from the unthresholded detections in Table 4.7 to the thresholded detections in Table 4.8 there are significant differences. Brwncamp1 Object 1, is an example where both the LW and MW detections increased significantly between unthresholded and thresholded data sets but the overall detection count only increased by 1 detection. On the other hand brwncamp1 Object 2 increased from 80 in LW and 75 in MW to 99 in LW and 99 in MW. In 36 cases MWIR detected more frequently than LWIR. I wont list all cases here but in vons1 object 1,2, and 3 all were detected more frequently in MWIR than LWIR. It is unclear to me if this is because of ambient conditions while capturing data or other factors that improved MWIR performance. In vons6 objects 1, 2, 3, 4 and 7 the opposite is true.

Tables 4.10 and 4.9 aggregate the IoU > 0.7330 detections for thresholded and unthresholded detections. There are no classes in which YOLOv4 applied to MWIR outperformed LWIR on the unthresholded data set. However, after preprocessing MWIR outperformed LWIR in 3 classes. As we can

Table 4.7: Unthresholded YOLOv4 Detection Results (IoU > 0.7330)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp1	1	SUV	48	41	53	9.43
brwncamp1	2	Car	117	7	118	0.85
brwncamp1	4	Car	80	75	94	14.89
brwncamp1	6	Car	17	1	18	5.56
brwncamp1	7	Car	217	87	234	7.26
brwncamp1	8	Car	76	34	79	3.80
brwncamp1	9	SUV	35	10	40	12.50
brwncamp1	10	Car	125	48	133	6.02
brwncamp1	11	Pickup	57	17	60	5.00
brwncamp1	12	Car	63	27	68	7.35
brwncamp2	5	Pickup	102	44	121	15.70
brwncamp2	7	Pickkup & Trailer	50	3	53	5.66
brwncamp2	8	Pickup	64	45	85	24.71
brwncamp2	9	Car	53	17	56	5.36
brwncamp2	10	Car	106	98	121	12.40
brwncamp2	11	SUV	93	53	118	21.19
brwncamp3	2	Pickup	131	96	137	4.38
brwncamp3	3	Van	141	106	143	1.40
brwncamp3	4	Car	58	68	91	36.26
brwncamp3	5	Pickup & Trailer	26	26	27	3.70
brwncamp3	6	Pickup	7	2	8	12.50
brwncamp3	7	SUV & Trailer	96	77	108	11.11
brwncamp3	8	SUV	99	83	113	12.39
brwncamp3	9	Car	115	124	144	20.14
brwncamp4	2	Semi & Trailer	1	1	2	50.00
brwncamp4	3	Semi & Trailer	25	25	33	24.24
brwncamp4	5	Car	3	2	3	0.00
brwncamp4	8	Car	2	4	4	50.00
brwncamp4	9	Van	6	5	8	25.00
brwncamp4	10	Pickup	26	26	29	10.34
brwncamp4	11	Car	75	24	76	1.32

Table 4.7: Unthresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp5	1	Pickup	73	28	82	10.98
brwncamp5	2	Car	126	132	167	24.55
brwncamp5	3	Car	127	119	174	27.01
brwncamp5	4	SUV	10	8	18	44.44
brwncamp5	6	Pickup	87	30	91	4.40
brwncamp5	7	Box Truck	34	2	36	5.56
brwncamp5	8	SUV & Trailer	116	63	125	7.20
brwncamp5	9	SUV & Trailer	87	36	92	5.43
brwncamp5	10	Semi	146	99	183	20.22
brwncamp5	11	Pickup	79	51	80	1.25
brwncamp5	12	Pickup	9	1	10	10.00
brwncamp5	13	Car	5	4	8	37.50
brwncamp5	14	Car	91	38	92	1.09
brwncamp5	15	Pickup	44	16	44	0.00
brwncamp5	18	Pickup	23	3	23	0.00
brwncamp6	1	Car	51	11	55	7.27
brwncamp6	3	Car	23	2	23	0.00
brwncamp6	7	Car	15	1	16	6.25
brwncamp6	8	SUV	78	24	81	3.70
brwncamp6	9	Pickup	100	6	100	0.00
brwncamp6	10	SUV	169	75	177	4.52
brwncamp6	11	Pickup	67	2	69	2.90
brwncamp6	13	SUV	47	5	47	0.00
brwncamp6	14	Car	2	11	13	84.62
brwncamp6	15	Van	60	8	60	0.00
brwncamp7	1	SUV	53	41	53	0.00
brwncamp7	2	SUV	72	66	74	2.70
brwncamp7	3	Pickup	105	39	111	5.41
brwncamp7	4	Car	26	6	26	0.00
brwncamp7	5	Car	64	49	66	3.03
brwncamp7	6	Car	104	94	113	7.96
brwncamp7	7	Pickup	89	97	108	17.59
brwncamp7	9	Box Truck	122	25	129	5.43
brwncamp7	10	Pickup	93	85	103	9.71

Table 4.7: Unthresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp
brwncamp7	11	Car	66	82	91	27.47
brwncamp7	12	Pickup	111	91	117	5.13
brwncamp7	13	Car	88	8	92	4.35
brwncamp7	14	Pickup	200	130	217	7.83
brwncamp7	15	Pickup	40	23	56	28.57
brwncamp7	17	Pickup?	2	1	3	33.33
brwncamp8	1	Person	774	428	811	4.56
brwncamp8	2	Person	312	130	339	7.96
brwncamp8	4	Pickup	70	33	73	4.11
brwncamp8	6	SUV	61	2	63	3.17
brwncamp8	7	SUV	71	4	72	1.39
brwncamp8	8	Pickup	72	17	74	2.70
brwncamp8	9	Car	100	90	122	18.03
brwncamp8	10	SUV	34	3	36	5.56
brwncamp8	11	Pickup	83	10	89	6.74
brwncamp8	12	Pickup	16	3	17	5.88
brwncamp8	13	Pickup	94	21	99	5.05
brwncamp8	14	Car	98	3	98	0.00
brwncamp8	15	Pickup	83	1	84	1.19
brwncamp8	16	Pickup	78	3	80	2.50
brwncamp8	17	Car	36	3	39	7.69
brwncamp8	18	Car	42	20	58	27.59
brwncamp8	19	Pickup	76	1	77	1.30
brwncamp8	20	Pickup	42	1	42	0.00
brwncamp9	1	Pickup	6	1	7	14.29
brwncamp9	3	SUV	3	1	3	0.00
brwncamp9	5	Semi & Trailer	22	19	23	4.35
brwncamp9	6	SUV	4	1	4	0.00
brwncamp9	8	Car	3	3	5	40.00
brwncamp9	9	Car	4	3	5	20.00
brwncamp9	10	Car	5	2	7	28.57
brwncamp9	11	SUV	3	2	3	0.00
brwncamp9	12	Pickup	3	1	4	25.00
brwncamp9	13	Pickup	3	1	3	0.00
brwncamp9	14	Car	11	2	11	0.00

Table 4.7: Unthresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
SBAP12	2	Car	1	1	2	50.00
SBAP13	3	Car	51	1	51	0.00
SBAP17	3	SUV	83	12	83	0.00
SBAP19	1	Airplane	6	2	7	14.29
SBAP24	1	Fuel Truck	421	1	421	0.00
SBAP27	1	Fuel Truck	142	31	144	1.39
SBAP5	2	Fuel Truck	6	4	10	40.00
vons1	1	Person	534	581	608	12.17
vons1	2	Person	24	53	61	60.66
vons1	3	Person	3	7	9	66.67
vons1	7	Pickup	1	1	1	0.00
vons2	2	SUV	70	36	77	9.09
vons3	1	Person	4	11	15	73.33
vons3	2	Car	52	16	52	0.00
vons4	1	Person	262	223	297	11.78
vons4	3	Person	17	3	19	10.53
vons5	1	Person	14	15	21	33.33
vons5	2	SUV	445	362	445	0.00
vons5	3	SUV	277	199	277	0.00
vons6	1	Person	1,054	607	1,066	1.13
vons6	2	Person	1,172	779	1,183	0.93
vons6	4	Person	205	127	205	0.00
vons6	7	Person	31	10	32	3.13
Sum	**All**		**12,102**	**6,779**	**13,036**	**7.71%**

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp1	1	SUV	53	53	54	1.89
brwncamp1	2	Car	99	21	107	8.08
brwncamp1	3	Car	100	7	103	3.00
brwncamp1	4	Car	82	78	94	14.63
brwncamp1	5	Car	117	1	118	0.85
brwncamp1	6	Car	23	1	24	4.35
brwncamp1	7	Car	218	98	234	7.34
brwncamp1	8	Car	83	42	84	1.20
brwncamp1	9	SUV	44	11	50	13.64
brwncamp1	10	Car	121	51	127	4.96
brwncamp1	11	Pickup	56	17	59	5.36
brwncamp1	12	Car	65	27	70	7.69

see in Table 4.10 those classes had fewer frames than object classes where LWIR detected more. There were two box trucks in the data set and three pickups with trailers. There were a total of 7 distinct indeterminate detections after thresholding and none without pre-processing. There were a total of eight indeterminate objects, with 3,451 labelled bounding boxes. The maximum size of indeterminate objects in terms of pixel area is 1,131 pixels and the minimum size is 10 pixels. There were two conditions that made determining the object challenging, the first was small and distant objects. The second was fast moving object close the camera system. This second class of object could exceed the cameras field-of-view and detecting on partial objects with motion blur proved to be challenging for both the labellers and the CNNs.

Table 4.12 and Table 4.13 are the improvement in detections between LW and MW and DB relative to the IoU threshold. These values were calculated by taking the unique ground truth object which had a detection with an

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp2	1	SUV	22	0	22	0.00
brwncamp2	2	Pickup	30	0	30	0.00
brwncamp2	3	SUV & Trailer	41	5	44	7.32
brwncamp2	4	Pickup & Trailer	6	0	6	0.00
brwncamp2	5	Pickup	94	149	175	17.45
brwncamp2	6	Car	30	32	56	75.00
brwncamp2	7	Pickup & Trailer	26	107	123	14.95
brwncamp2	8	Pickup	41	135	148	9.63
brwncamp2	9	Car	68	53	79	16.18
brwncamp2	10	Car	103	103	117	13.59
brwncamp2	11	SUV	75	112	149	33.04
brwncamp2	12	Pickup	25	0	25	0.00
brwncamp3	1	Car	1	0	1	0.00
brwncamp3	2	Pickup	138	98	140	1.45
brwncamp3	3	Van	148	138	166	12.16
brwncamp3	4	Car	71	94	108	14.89
brwncamp3	5	Pickup & Trailer	25	25	25	0.00
brwncamp3	6	Pickup	17	13	21	23.53
brwncamp3	7	SUV & Trailer	105	114	121	6.14
brwncamp3	8	SUV	102	112	117	4.46
brwncamp3	9	Car	135	137	159	16.06
brwncamp3	10	Car	43	0	43	0.00
brwncamp3	11	Indeterminate	1	0	1	0.00
brwncamp4	2	Semi & Trailer	3	3	4	33.33
brwncamp4	3	Semi & Trailer	23	26	33	26.92
brwncamp4	5	Car	3	2	3	0.00
brwncamp4	8	Car	2	4	4	0.00
brwncamp4	9	Van	6	5	8	33.33
brwncamp4	10	Pickup	26	26	29	11.54
brwncamp4	11	Car	75	24	76	1.33
brwncamp4	12	Indeterminate	26	0	26	0.00
brwncamp4	13	Indeterminate	8	0	8	0.00

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp5	1	Pickup	70	71	96	35.21
brwncamp5	2	Car	142	167	185	10.78
brwncamp5	3	Car	153	204	245	20.10
brwncamp5	4	SUV	12	40	46	15.00
brwncamp5	6	Pickup	92	52	100	8.70
brwncamp5	7	Box Truck	23	56	62	10.71
brwncamp5	8	SUV & Trailer	131	139	162	16.55
brwncamp5	9	SUV & Trailer	65	42	75	15.38
brwncamp5	10	Semi	210	206	239	13.81
brwncamp5	11	Pickup	94	88	103	9.57
brwncamp5	12	Pickup	8	27	34	25.93
brwncamp5	13	Car	9	19	25	31.58
brwncamp5	14	Car	96	87	102	6.25
brwncamp5	15	Pickup	58	59	76	28.81
brwncamp5	16	Indeterminate	19	0	19	0.00
brwncamp5	17	Van	35	5	38	8.57
brwncamp5	18	Pickup	33	31	42	27.27
brwncamp6	1	Car	66	28	71	7.58
brwncamp6	2	Semi & Trailer	39	0	39	0.00
brwncamp6	3	Car	48	1	48	0.00
brwncamp6	4	Car	4	2	6	50.00
brwncamp6	7	Car	13	1	14	7.69
brwncamp6	8	SUV	87	31	95	9.20
brwncamp6	9	Pickup	108	18	114	5.56
brwncamp6	10	SUV	220	55	226	2.73
brwncamp6	11	Pickup	41	3	43	4.88
brwncamp6	12	SUV	52	0	52	0.00
brwncamp6	13	SUV	53	6	54	1.89
brwncamp6	14	Car	9	27	32	18.52
brwncamp6	15	Van	64	10	64	0.00

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp7	1	SUV	53	44	53	0.00
brwncamp7	2	SUV	75	68	75	0.00
brwncamp7	3	Pickup	108	58	121	12.04
brwncamp7	4	Car	31	10	31	0.00
brwncamp7	5	Car	64	65	74	13.85
brwncamp7	6	Car	114	105	129	13.16
brwncamp7	7	Pickup	89	108	116	7.41
brwncamp7	8	Indeterminate	6	0	6	0.00
brwncamp7	9	Box Truck	127	121	156	22.83
brwncamp7	10	Pickup	94	92	108	14.89
brwncamp7	11	Car	63	85	93	9.41
brwncamp7	12	Pickup	133	150	171	14.00
brwncamp7	13	Car	87	36	105	20.69
brwncamp7	14	Pickup	211	218	242	11.01
brwncamp7	15	Pickup	62	68	101	48.53
brwncamp7	16	Pickup	1	1	2	100.00
brwncamp7	17	Pickup?	13	4	14	7.69
brwncamp8	1	Person	810	588	877	8.27
brwncamp8	2	Person	438	250	477	8.90
brwncamp8	3	Indeterminate	6	1	7	16.67
brwncamp8	4	Pickup	98	49	104	6.12
brwncamp8	6	SUV	65	1	66	1.54
brwncamp8	7	SUV	79	61	93	17.72
brwncamp8	8	Pickup	88	89	106	19.10
brwncamp8	9	Car	130	158	184	16.46
brwncamp8	10	SUV	52	50	92	76.92
brwncamp8	11	Pickup	95	43	113	18.95
brwncamp8	12	Pickup	17	5	18	5.88
brwncamp8	13	Pickup	106	44	121	14.15
brwncamp8	14	Car	118	18	122	3.39
brwncamp8	15	Pickup	91	12	95	4.40
brwncamp8	16	Pickup	109	9	114	4.59
brwncamp8	17	Car	52	63	103	63.49
brwncamp8	18	Car	65	87	121	39.08
brwncamp8	19	Pickup	79	63	126	59.49
brwncamp8	20	Pickup	39	0	39	0.00

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
brwncamp9	1	Pickup	5	2	7	40.00
brwncamp9	2	Pickup	6	0	6	0.00
brwncamp9	3	SUV	3	1	3	0.00
brwncamp9	5	Semi & Trailer	22	18	23	4.55
brwncamp9	6	SUV	4	1	4	0.00
brwncamp9	7	Car	1	0	1	0.00
brwncamp9	8	Car	3	3	5	66.67
brwncamp9	9	Car	4	3	5	25.00
brwncamp9	10	Car	5	2	7	40.00
brwncamp9	11	SUV	3	2	3	0.00
brwncamp9	12	Pickup	3	1	4	33.33
brwncamp9	13	Pickup	3	1	3	0.00
brwncamp9	14	Car	12	2	12	0.00
brwncamp9	15	Car	2	0	2	0.00
brwncamp9	16	Car	1	0	1	0.00
brwncamp9	17	Car	3	0	3	0.00
SBAP10	2	Bird	1	0	1	0.00
SBAP10	3	Pickup	19	0	19	0.00
SBAP10	7	Car	17	0	17	0.00
SBAP10	8	Bird	1	0	1	0.00
SBAP10	10	Pickup	1	0	1	0.00
SBAP10	11	Car	20	0	20	0.00
SBAP12	1	Pickup	27	0	27	0.00
SBAP12	2	Car	3	2	5	66.67
SBAP13	1	Car	26	0	26	0.00
SBAP13	3	Car	56	8	56	0.00
SBAP13	4	Bird	1	1	1	0.00
SBAP14	1	Airplane	272	0	272	0.00
SBAP15	1	Airplane	298	15	298	0.00
SBAP16	1	SUV	19	0	19	0.00
SBAP16	2	Car	12	0	12	0.00
SBAP16	3	SUV	12	1	13	8.33
SBAP17	2	Car	19	0	19	0.00
SBAP17	3	SUV	82	15	82	0.00
SBAP17	4	SUV	43	4	43	0.00
SBAP17	6	Pickup	4	0	4	0.00

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
SBAP18	1	Fuel Truck	4	0	4	0.00
SBAP18	2	SUV	8	0	8	0.00
SBAP18	3	Car	20	1	20	0.00
SBAP19	1	Airplane	7	4	10	42.86
SBAP2	1	Airplane	21	0	21	0.00
SBAP2	5	Car	1	0	1	0.00
SBAP20	1	Airplane	143	0	143	0.00
SBAP21	2	Airplane	180	0	180	0.00
SBAP22	2	Airplane	149	0	149	0.00
SBAP22	3	Car	34	2	34	0.00
SBAP23	1	Airplane	78	0	78	0.00
SBAP24	1	Fuel Truck	454	85	468	3.08
SBAP25	1	Car	24	0	24	0.00
SBAP25	2	Airplane	162	2	162	0.00
SBAP25	3	Airplane	1	0	1	0.00
SBAP26	1	Airplane	66	1	67	1.52
SBAP27	1	Fuel Truck	149	85	169	13.42
SBAP3	1	Fuel Truck	52	34	72	38.46
SBAP5	1	Airplane	2	0	2	0.00
SBAP5	2	Fuel Truck	9	8	12	33.33
SBAP6	1	Airplane	1	0	1	0.00
SBAP7	1	Airplane	52	0	52	0.00
SBAP8	2	Bird	9	0	9	0.00
SBAP9	1	Airplane	417	0	417	0.00

Table 4.8: Thresholded YOLOv4 Detection Results (IoU > 0.7330) (Cont.)

Sequence	Obj.	Class	LW Num	MW Num	DB Num	%-Imp.
vons1	1	Person	534	581	608	4.65
vons1	2	Person	24	53	61	15.09
vons1	3	Person	3	7	9	28.57
vons1	7	Pickup	1	1	1	0.00
vons2	2	SUV	70	36	77	10.00
vons3	1	Person	6	11	17	54.55
vons3	2	Car	51	38	51	0.00
vons4	1	Person	275	223	309	12.36
vons4	3	Person	17	3	19	11.76
vons5	1	Person	14	15	21	40.00
vons5	2	SUV	446	380	446	0.00
vons5	3	SUV	277	195	277	0.00
vons6	1	Person	1054	607	1066	1.14
vons6	2	Person	1172	779	1183	0.94
vons6	3	Person	17	0	17	0.00
vons6	4	Person	205	127	205	0.00
vons6	7	Person	31	10	32	3.23
Sum	**All**		**12,432**	**9,325**	**14,414**	**14.74%**

Table 4.9: Unthresholded YOLO v4 Detection Results by Object Class (IoU > 0.7330)

Class	LW Num	MW Num	DB Num	%-Improved
Airplane	6	2	7	14.29
Box Truck	156	27	165	5.56
Car	2,298	1,317	2,625	7.27
Fuel Truck	569	36	575	40.00
Person	4,406	2,974	4,666	4.56
Pickup	2,134	926	2,301	0.00
Pickup & Trailer	76	29	80	5.66
Pickup?	2	1	3	33.33
Semi	146	99	183	20.22
Semi & Trailer	48	45	58	50.00
SUV	1,755	1,028	1,837	3.70
SUV & Trailer	299	176	325	7.20
Van	207	119	211	0.00

Table 4.10: Thresholded YOLOv4 Detection Results by Object Class (IoU > 0.7330)

Class	LW Num	MW Num	DB Num	%-Improved
Box Truck	150	177	218	31.19
Car	2,734	1,986	3,333	17.97
Indeterminate	6	1	7	14.29
Person	4,583	3,254	4,884	6.16
Pickup	2,269	1,801	2,853	20.47
Pickup & Trailer	51	132	148	65.54
Pickup?	13	4	14	7.14
Semi	210	206	239	12.13
Semi & Trailer	48	47	60	20.00
SUV	1,773	1,259	1,980	10.45
SUV & Trailer	342	300	402	14.93
Van	253	158	276	8.33

associated IoU that exceeds the threshold. The "%-Improved" column shows the percent improvement over the greatest of the LW and MW column, in all cases it was LW. The percent improvement was calculated as

$$\%-improvement = \frac{DB\# - LW\#}{LW\#}. \tag{4.5}$$

Where # represents the count of the unique ground truth. One of the central points of this dissertation is that the addition of a second infrared spectral band improves the performance of CNN based object detectors. In the case of YOLOv4 that amount exceeds an 8% improvement. Detections with a class id of 0 which corresponds to background were excluded for YOLOv4 however YOLOv7 does not have a background class ID. The bold elements of Table 4.12 and Table 4.13 represent the higher value between the two tables. In other cases classification was ignored as this dissertation focuses primarily detection and not classification.

Table 4.11 shows the additional detections in the sense of TP that results from including Background Class. However, what is not shown here is the number of FP and the resultant degradation of precision. That is, the recall number increases but the precision is adversely affected. To examine the precision and recall the values were calculated according to Eq. 2.1 and Eq. 2.2 respectively for all values of CNN classification probability from 0.05 to 0.95 in 0.05 increments. The results for YOLOv4 are displayed in Figs. 4.16, 4.17, and 4.18 for IoU values of 0.25, 0.50, and 0.75. The general interpretation of these plots is that the MW has very high precision compared to LW but very low recall in comparison. Given the introduction of detections from the MW that are not present in the LW, the DB has the dual benefit of adding new

Table 4.11: YOLOv4 Detection Improvement per IoU Threshold
All Classes

IoU	LW	MW	DB	%-Improved
0.95	2,658	1,160	3,287	23.66
0.9	7,251	3,936	8,458	16.64
0.8	14,395	8,285	15,928	10.64
0.7	18,603	10,704	20,501	10.20
0.6	20,958	12,100	23,105	10.24
0.5	22,616	12,932	24,920	10.18
0.4	24,265	13,712	26,743	10.21
0.3	26,025	14,640	28,683	10.21
0.2	27,638	15,786	30,568	10.60
0.1	29,749	17,533	32,931	10.69
0	35,838	35,449	47,257	31.86

detections that have a low FN rate. The effect of increasing IoU is that the recall rate is diminished as IoU increases. To get a better understanding of IoU impact on precision and recall, Fig. 4.25 shows the curves for IoU varying form 0.05 to 0.95 in 0.05 increments. Comparing the plots in Fig. 4.25 and Figs. 4.16, 4.16, and 4.18 to Table 4.12 and Table 4.13 the position of the green DB are to the right, i.e. have higher recall than the red LW and black DB. The percentage to the right of LW which is greater than MW in all cases is the percent improvements from these tables. The black dotted lines were added to Figs. 4.16, 4.16, and 4.18 to allow the reader to visually connect the points on the MW curve to the corresponding points on the DB curve. The circles in these plots allow the reader to compare the relationship of the recall between the LW and DB curves.

Table 4.12 and Table 4.13 show TP for IoU ranging from 0 to 0.95. The reason it is important to understand this is to show that the benefit of using

Table 4.12: YOLOv4 Detection Improvement per IoU Threshold - Background Removed

IoU	LW	MW	DB	%-Improved
0.95	**2,388**	799	**2,847**	19.22
0.9	**6,758**	2,830	**7,646**	13.13
0.8	13,518	6,222	14,651	8.38
0.7	17,395	8,163	18,788	8.00
0.6	19,548	9,339	21,174	8.31
0.5	21,071	10,020	22,830	8.34
0.4	22,560	10,615	24,444	8.35
0.3	24,189	11,276	26,190	8.27
0.2	25,749	12,036	27,893	8.32
0.1	27,804	13,351	30,110	8.29
0	33,099	**27,886**	**42,955**	29.77

Table 4.13: YOLOv7 Detection Improvement per IoU Threshold

IoU	LW	MW	DB	%-Improved
0.95	2,136	**1,031**	2,553	19.52
0.9	5,772	**3,470**	6,783	17.51
0.8	**13,369**	**9,137**	**15,147**	13.29
0.7	**18,780**	**13,160**	**21,038**	12.02
0.6	**22,044**	**15,435**	**24,479**	11.04
0.5	**24,553**	**16,731**	**27,019**	10.04
0.4	**26,267**	**17,562**	**28,865**	9.89
0.3	**27,520**	**18,480**	**30,247**	9.90
0.2	**29,111**	**19,451**	**31,961**	9.79
0.1	**30,876**	**20,978**	**33,967**	10.01
0	**33,721**	24,329	37,695	11.78

DBIR is not dependent on an arbitrarily selected value. **These two tables show that the benefit is robust against IoU selection.**

In the case of YOLOv7, the improvement exceeds 9.79%. If no class filtering is used overall YOLOv4 detected **26.12%** of the labeled ground truth at 0.7330 IoU, compared to YOLOv7, which detected **28.55%** of labeled ground truth. Removing Class 0 reduced YOLOv4 percent detected to **22.01%** at 0.7330 IoU. In this case, the detection rate as a percent is marginally better for YOLOv7 compared to YOLOv4. Unfortunately, it is difficult to determine if this marginal improvement in TP performance is consistent with the authors of [186], because the only comparison made in that paper to YOLOv4 is a speed test with a scaled version of YOLOv4.

To understand the relationship between these percentages, which are the recall values at 0.7330 IoU compare to Fig. 4.18 and Fig. 4.21. When comparing the performance of YOLOv4 to YOLOv7 on this dataset, it is shown in Fig. 4.18 and Fig. 4.21 that YOLOv7 substantially increased the recall in both the MW and LW compared to YOLOv4, with MW being a more significant percent increase. Along with the increase in TP, there is a noteworthy increase in FP as well. The addition in FP in MW contributes the rate of FP in DB. In this case DB performs poorer than LW in terms of precision, i.e. has a higher FP rate. However, TP values are still significantly higher in DB than either MW or LW, i.e a much higher precision for DB.

In Fig. 4.25 and Fig. 4.26 precision-recall curves for each IoU value are plotted. Taking the maximum value across all confidences per curve, Fig. 4.27 and Fig. 4.27 were generated. Assuming an optimal confidence threshold was selected Fig. 4.27 and Fig. 4.28 illustrate that the best achievable precision-

recall performance on this dataset for each IoU. In Fig. 4.27(a), it can be observed that the DB benefits from the high recall of LW and the additional recall in MW. The extremely high precision of MW in Fig. 4.27(a) indicates that DB doesn't suffer a precision penalty due to the combination of MW and LW for YOLOv4. However, YOLOv7 differs on this data set in the MW the recall is nearly double of the MW recall for YOLOv4 but the precision is significantly lowered. The recall for all curves are higher for YOLOv7. The additional FP stemming from the decreased MW precision increases the overall FP rate, lowering the precision of DB relative to LW.

For certain confidence-IoU pairs, YOLOv7 produced very few or even no detections. This resulted in the points for the curve to be removed to avoid divide by zero, or a statistically dubious value for precision output. It should not be interpreted that YOLOv7 did exceptionally well but is an artifact of small numbers and division. This phenomenon is apparent at IoU of 0.9 in Fig. 4.28(b).

What can be concluded from Fig. 4.27 is that doing inference on DBIR data improves the recall significantly for all values of IoU when choosing an optimal confidence threshold. Similarly from Fig. 4.27 we can conclude that we get greater improvement to recall but we do take a hit to precision. Techniques for mitigating False Positive (FP) (FP) were discussed in the Chapter 2 of this dissertation so the impact can likely be mitigated making the trade-off justifiable. Further research into applying these techniques to this dataset and the output of CNNs on it would likely be beneficial.

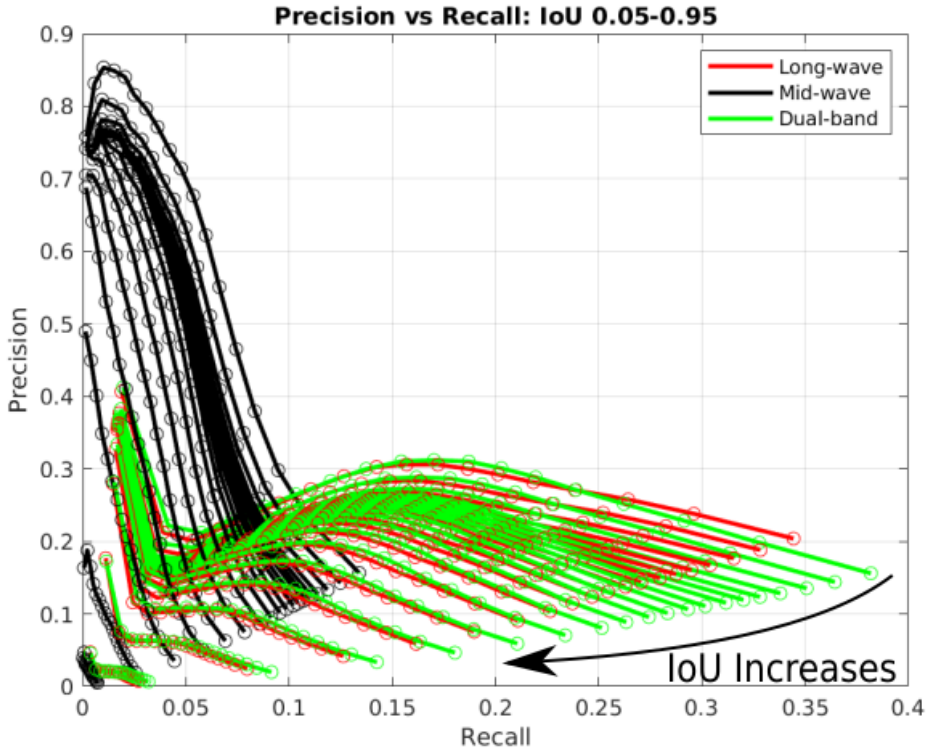


Figure 4.25: YOLOv4 LW, MW, and DB for IoU 0.05 to 0.95 with 0.05 increment.

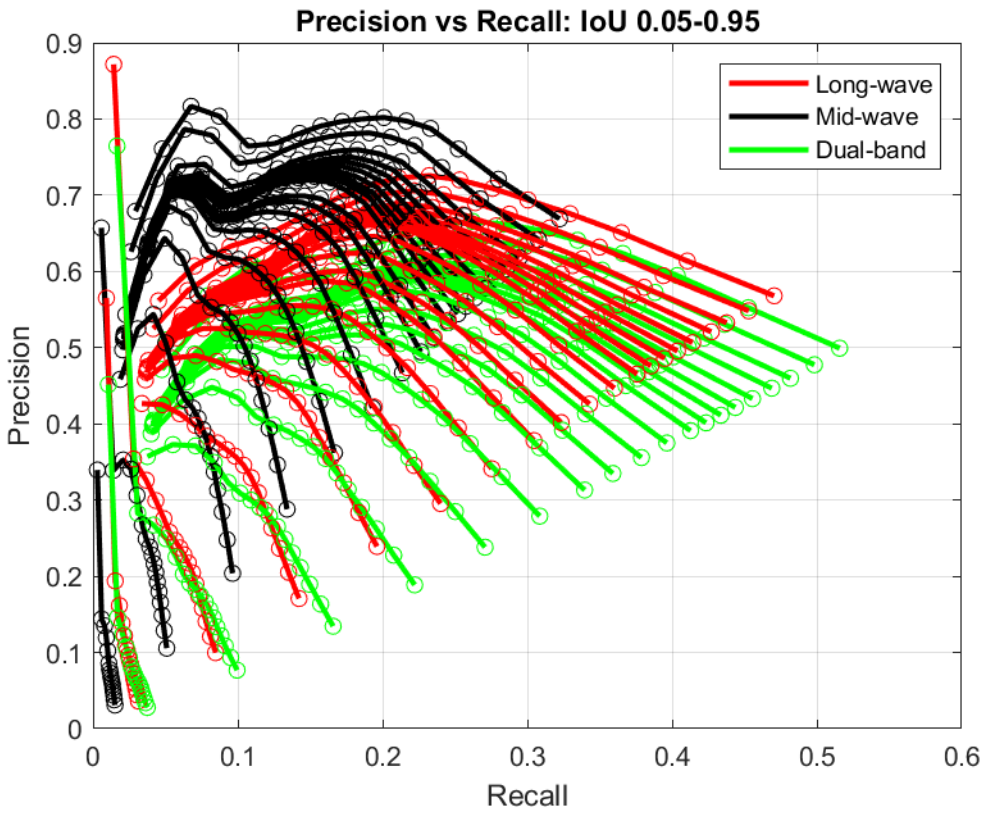
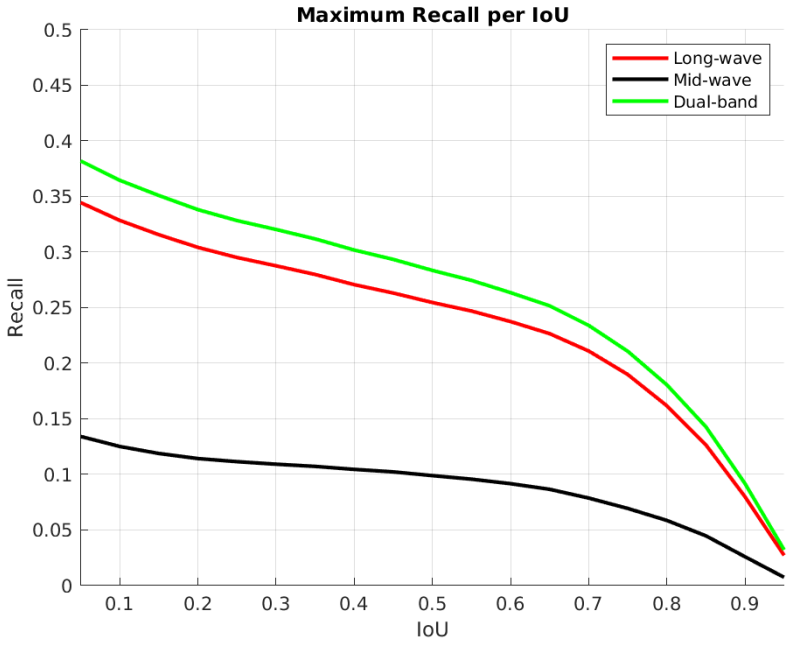
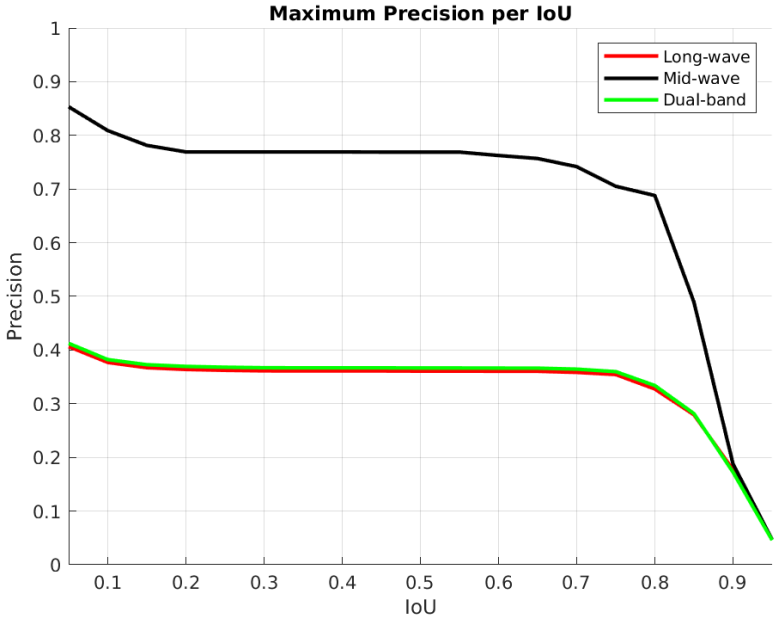


Figure 4.26: YOLOv7 LW, MW, and DB for IoU 0.05 to 0.95 with 0.05 increment.

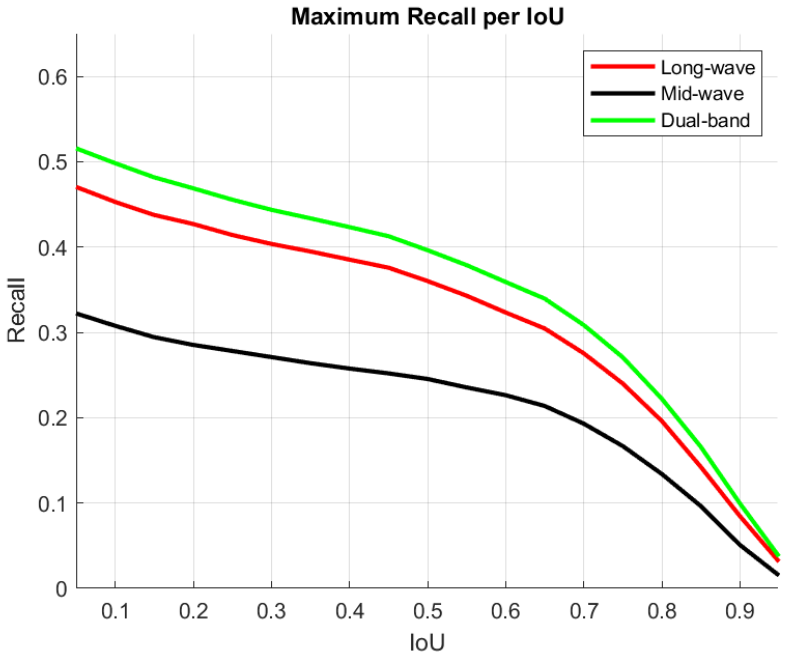


(a) YOLOv4 Maximal Recall per IoU

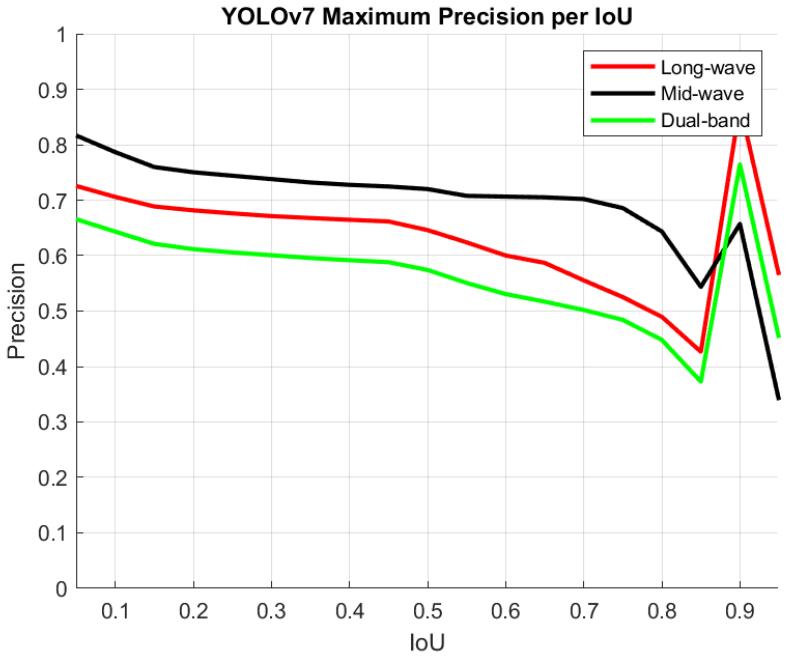


(b) YOLOv4 Maximal Precision per IoU

Figure 4.27: YOLOv4 Maximum Precision and Recall per IoU



(a) YOLOv7 Maximal Recall per IoU



(b) YOLOv7 Maximal Precision per IoU

Figure 4.28: YOLOv7 Maximum Precision and Recall per IoU

The approach to counting FP is that every occurrence is individually counted. CNNs often yield miltiple detections at the same spot and report different classes and class-probabilities. These detections often disagree, especially in YOLO, by fractions of pixels in x, y, height, and width. To reduce that number it is possible to cluster bounding boxes of FP to reduce the number of FP reporting the same location. This can be accomplished through Non-Maxima Suppression as discussed in Chapter 2. It is likely that the threshold would be fairly high and only reduce the count of FP that had extremely similar bounding-boxes. This was not done on any of the results here but in terms of metrics it is worthwhile considering if FP have been overcounted.

The greatest impact on detection performance of both YOLO variants tested here was the size of the object. This is apparent in the percent detected in Fig. 4.29 and Fig. 4.30. In Fig. 4.31 and Fig. 4.32 the raw TP/FP counts are given as stacked bar so that the reader can assess the statistical significance of the results reported in percentages. What we can see is that for YOLOv4 the detection performance picks up significantly for targets spanning greater than 30^2 pixels. The same is true for YOLOv7 but overall detection rate is slightly higher across all bounding box size categories. There is also a strong positive correlation in detection and object size, though the results appear non-linear as shown in the plots.

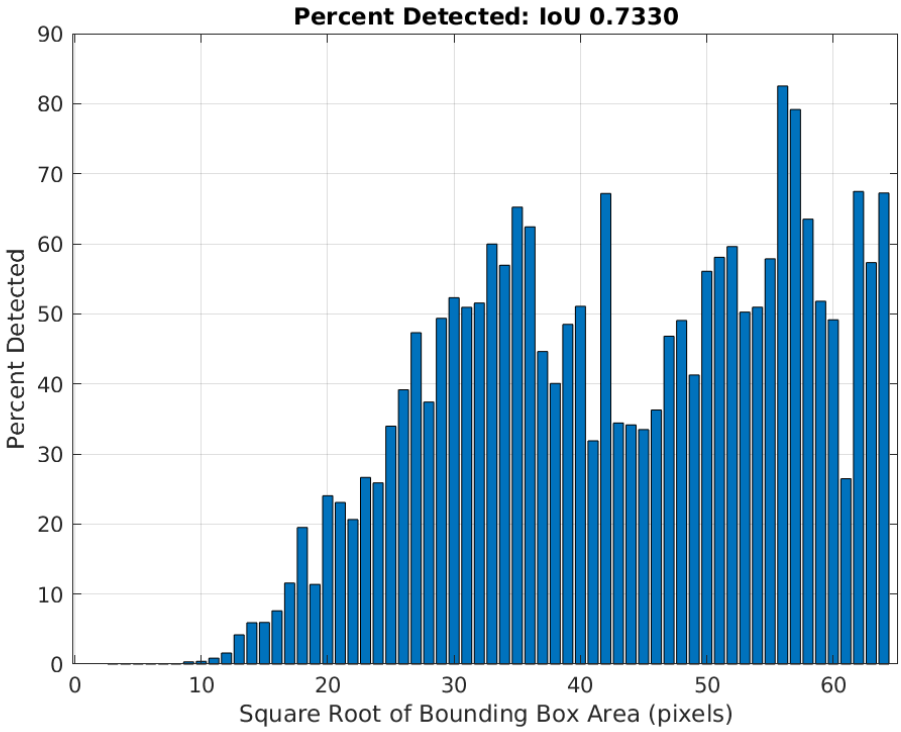


Figure 4.29: Percent of Object Detected by Bounding Box sizes. The rightmost column represents all objects over the size of 64×64 pixels.

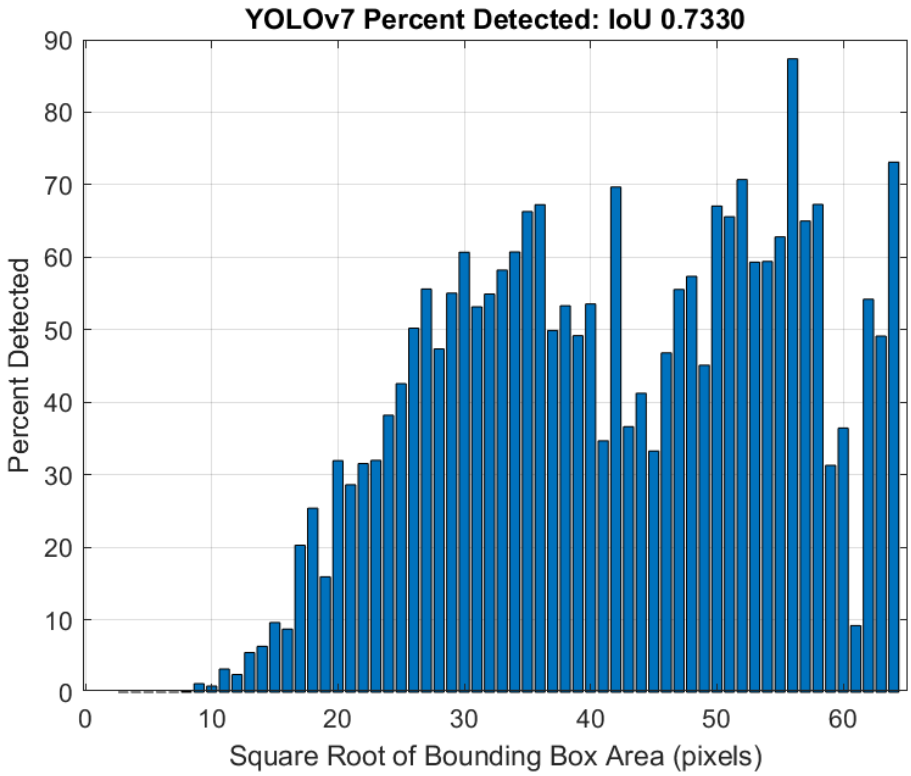


Figure 4.30: Percent of Object Detected by Bounding Box sizes. The rightmost column represents all objects over the size of 64×64 pixels.

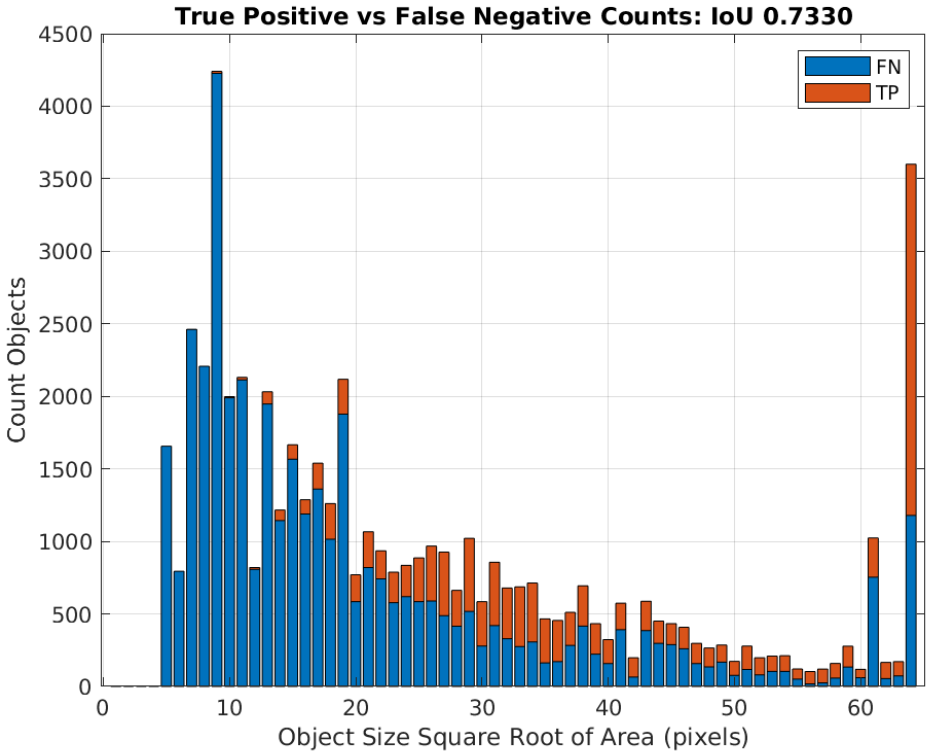


Figure 4.31: Number of objects detected by YOLOv4 (TP) stacked over number of objects withing the size class in the data set not detected (FN). The rightmost entry is all object 64^2 or greater.

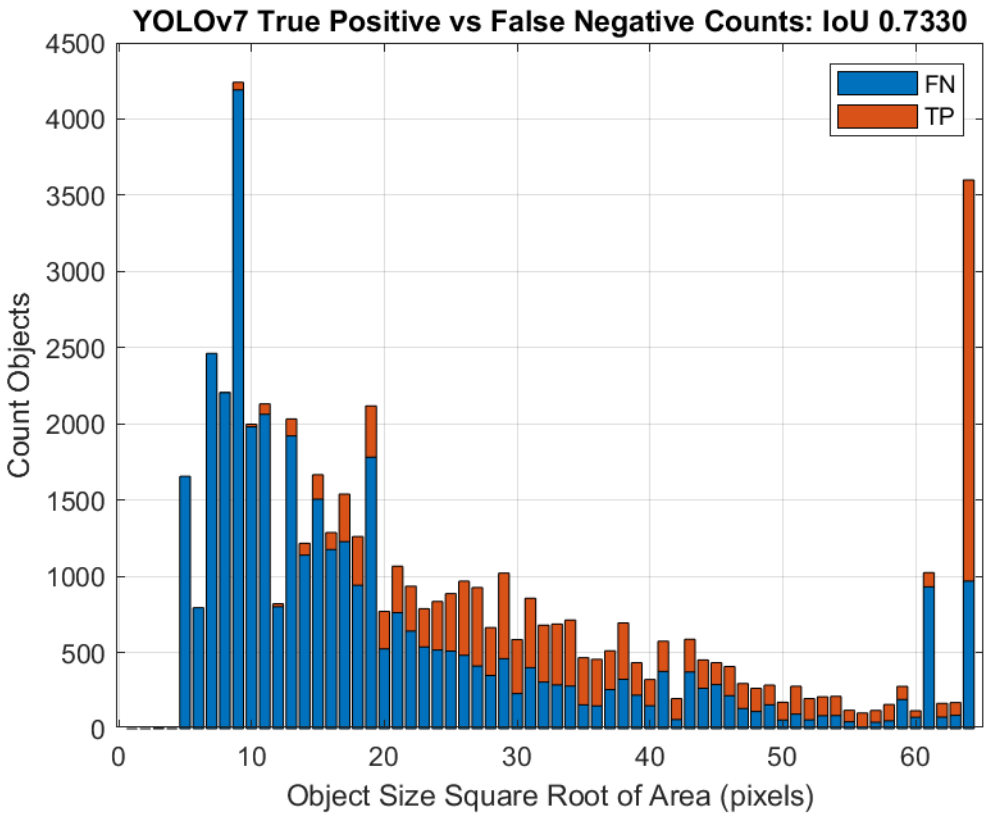


Figure 4.32: Number of objects detected by YOLOv7 (TP) stacked over number of objects withing the size class in the data set not detected (FN). The rightmost entry is all object 64^2 or greater.

4.1.5 Integration of YOLO and ViBe

It should be noted that the two detection approaches evaluated in this dissertation, viz. ViBe and YOLO, could be used in conjunction. A feasible system could be developed that combines both a CNN like YOLO and ViBe into a single surveillance system. One approach would be to employ one or more staring cameras with relatively wide FoV. These cameras could be designed to give wide coverage of the surveillance volumes. In such a setup, ViBe would be used to process the output from the staring camera to generate detections. A second camera system with a narrower field of view could be used with zoom capabilities and put more "pixels on target" using techniques such as those described in [41]. Then the camera system could track over detections with detections provided by a CNN variant like YOLO.

Another approach would be a stare then reposition algorithm. This still could be done with two camera systems but this time in the same Pan-Tilt Unit (PTU). In this way the two cameras systems could have mechanically registered parallel optical axes and could do both kinds of detections. If the PTU tracks the objects motion detection will be problematic. However, when in search mode the camera could be stationary for a sufficient duration to establish a background model and coduct motion detection when not actively tracking.

Fig. 4.33 and Fig. 4.34 stack TP detection counts for ViBe with an IoU of 0.5 and TP on YOLO detections at Iou 0.7330. In this way the complementary nature of size detection capabilities can visualized between YOLO and and algorithm like ViBe. ViBe is stronger at small object detection while the YOLO variants are stronger at Medium and Large Object detections.

Fig. 4.33 and Fig. 4.34 show that using a motion detection algorithm to detect small distant object then employing a camera system that can get a larger representation of the object are supportive of this hybridized approach. Several technical details would need to be managed in either of these approaches, if a significant portion of the motion detection pixels are activated that means that either the camera is moving or there is something in the near-FoV. ViBe does particularly well detecting smaller objects and systematically inspecting those smaller objects at a higher zoom level would be an effective approach to surveillance. This however doesn't mean that ViBe would not be aware the a significant portion of the FoV is in motion and as a suggestion tracking or clustering over ViBe detections may gather these bounding boxes into single objects and improve the large object performance. In this way the systems could be used to complement each other and increase the amount of track coverage and increase detection capability.

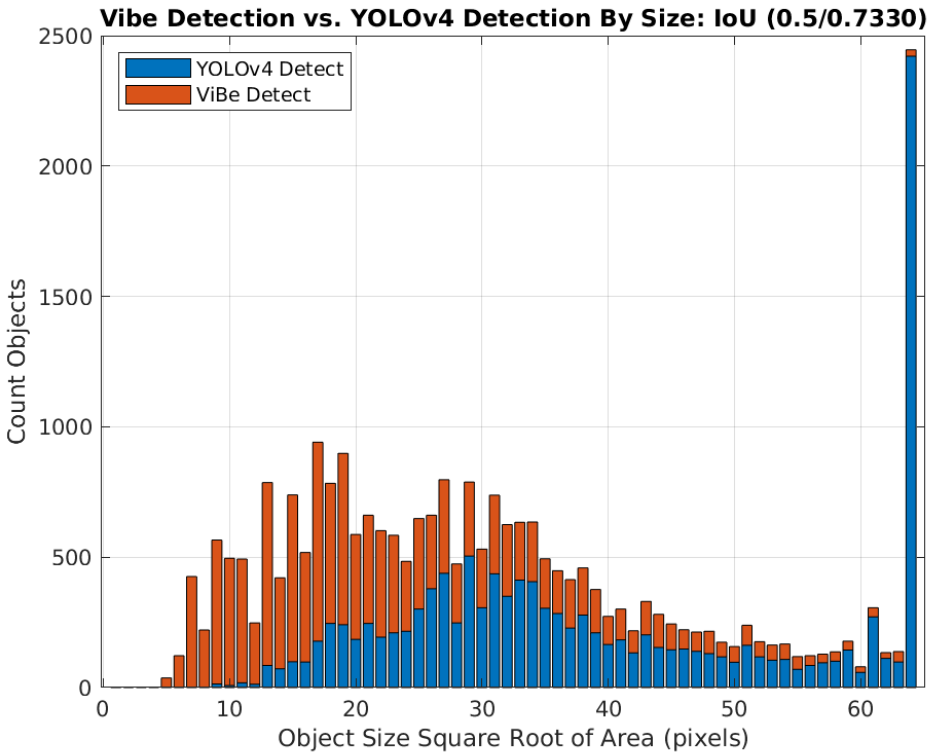


Figure 4.33: Detection capabilities of ViBe compared to YOLOv4 binned by the square-root of the object size. Note that the detection threshold for ViBe is 0.5 and YOLOv4 is 0.7330. The bin at x equals 64 contains all objects greater than 64^2 pixels in area.

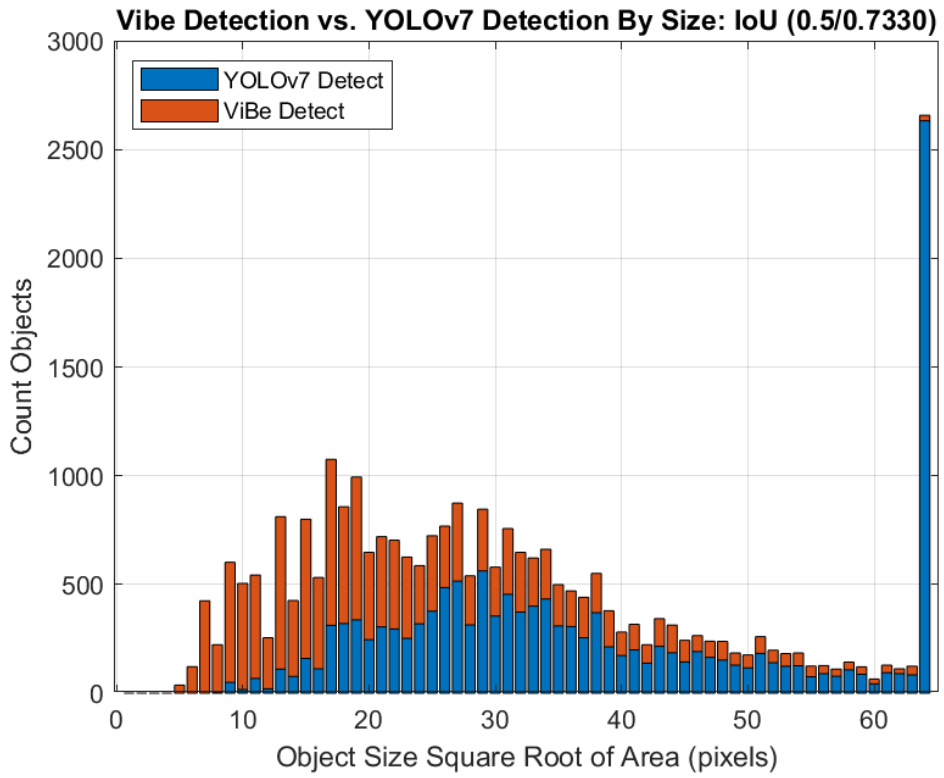


Figure 4.34: Detection capabilities of ViBe compared to YOLOv7 binned by the square-root of the object size. Note that the detection threshold for ViBe is 0.5 and YOLOv7 is 0.7330. The bin at x equals 64 contains all objects greater than 64^2 pixels in area.

Chapter 5

Conclusion

The original contributions of this dissertation are mainly in two areas of infrared computer vision. Firstly, a new labeled DBIR dataset is introduced. This dataset is comprehensively explained and demonstrated, distinguishing it from the only comparable dataset available. This new infrared data set contains over two hundred objects with over sixty thousand annotated bounding boxes. Additionally, it includes numerous long-track objects that can be utilized for researching target tracking algorithms.

Secondly, a preliminary evaluation for the benefits of DBIR over MWIR or LWIR data in terms of staet-of-the-arts motion detection and object detection techniques establishes the first evidence of the benefits of using DBIR in a contemporary context. These benefits are present in dataset without smoke or fog and extend the domain where an advantage was previously considered. This dataset serves as a testing ground for evaluating many other computer vision algorithms.

In the experimentation portion of this dissertation, the impact of using DBIR in one of the top-performing motion detection algorithms, ViBe, was evaluated. The results revealed that through a straightforward data fusion approach, combining MWIR and LWIR motion data contributed to larger and more coherent bounding boxes. This enhanced the localization and object size

qualities of detections with ViBe. Furthermore, it was established that ViBe excels at detecting small and distant objects in DBIR data. ViBe was able to detect objects that gave the human labellers tasked with creating the data set difficulty in detecting. However, the ViBe-based algorithm had issues with larger objects in the foreground due to fragmentation of the bounding boxes and has challenges with camera ego-motion in the absence of compensation techniques.

In a second set of experiments, this data set was evaluated with the modern object detection algorithms YOLOv4 and YOLOv7. A detailed explanation of image pre-processing was given to faciliate the reproduction of the results presented in Chapter 2. A comparative evaluation of the Precision-Recall curves was generated showing a significant increase in Recall using DBIR data over using MWIR and LWIR alone. Further there was an explanation of why this is the case. Those reasons are that in YOLOv4 the detriment to Precision was minimal, and in YOLOv7 that as MW detection was greatly increased so, similarly, were the FP rates. This lowered the Precision for YOLOv7 in our approach. A discussion about how to mitigate FP was given in Chapter 2. In my opinion the increase in Recall is significant and the decrease in Precision is manageable especially with known mitigation techniques.

Another contribution to both the motion detection and CNN based detection in DBIR was a detection size analysis. While it is well known that CNNs, such as YOLOv4 and YOLOv7 have difficulty detecting smaller objects, it is shown that ViBe-based motion detection does not. In this way motion detection and CNNs are complimentary when it comes to detection capabilities per object size. Schemes of using both of these systems are presented demon-

strate how these two algorithms might be used together. Combining motion and CNN object detection is a matter of prior art, but the fact that DBIR can potentially benefit both are new contributions shown here for the first time.

After a long period of relative inactivity in the open literature, the original results presented in this dissertation establish DBIR as an area still worthy of significant future research investment. The original work reported here suggests several promising areas for future research.

Image inpainting is the art of repairing damaged or deterioriated images [197]. Given the damaged nature of the MWIR FPA in this data set creating plausible repair to those regions could in theory make the data set more useful. Developing GANs, described as a nascent technique in [197], using information form the LWIR image to replace missing information caused by damaged pixels in the MWIR images would be a worthwhile endeavor.

In [190] some initial studies on apply GANs for the purposes of Super-Resolution was discussed. Building on that work with this data set could also yield useful knowledge.

Studies involving presenting MWIR and LWIR information fuses into a single image to humans were presented in some of the most important literature in DBIR computer vision [54, 55]. Expanding this research to include the data presented here could be another possible avenue of continued investigation.

A topic that I am interested in particularly is looking at the comparison of FP rejection methods. Clutter rejection, Non-Maximal Suppression, and Data Association of nearest neighbors all present as likely candidates to reduce FP and make sensor systems more reliable. Using characteristics specific to

DBIR, such as LWIR and MWIR enhanced features seem likely to improve data association in tracking algorithms. Improvements in the ability to associate data allow for more efficient FP rejection. A detailed discussion of approaches from the data association and tracking literature can be found in [8]. Non-maximum suppression is a common FP reduction algorithm employed in CNN and Deep Learning communities [58]. Evaluating the combination of the FP suppression techniques would be a worthy study.

One avenue of research that was not attempted in this dissertation was attempting to retrain or perform transfer learning to a CNN on this data set. The main reason was to avoid overfitting data on a small data set. Expanding this data set with more DBIR data or carefully constructing a data split may allow the researcher to perform transfer learning in a responsible manner. Further research into this would be warranted. Similarly retraining CNNs on the FLIR ADAS data set and then applying them to this data set might yield useful results.

Other, slower-than-real-time CNNs, like faster-RCNN, might be applied to this data set to establish maximal capabilities with state of the art given unlimited run time to give an upper bound on the detection capabilities when not controlled for run time.

Bibliography

[1] S. B. Achal, C. D. Anger, J. E. McFee, and R. W. Herring, "Detection of surface-laid mine fields in VNIR hyperspectral high-spatial-resolution data," in *Detection and Remediation Technologies for Mines and Minelike Targets IV*, A. C. Dubey, J. F. Harvey, J. T. Broach, and R. E. Dugan, Eds., vol. 3710, International Society for Optics and Photonics. SPIE, 1999, pp. 808 – 818. [Online]. Available: https://doi.org/10.1117/12.357103

[2] M. Ai, T. Liu, H. Ying, Z. Yuan, J. Wang, and Y. Shang, "A direct and robust method for ego-motion estimation," in *2021 China Automation Congress (CAC)*, 2021, pp. 1349–1353.

[3] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[4] A. Amankwah and C. Aldrich, "Multiresolution image registration using spatial mutual information," in *2012 Oceans*, 2012, pp. 1–4.

[5] F. R. Anggaraksa, M. T. Yuriawan, K. L. Firmansyah, L. Yulianti, and A. Izzuddin, "Thermal infrared tracking system with yolov4 algorithm and visual servoing method," in *2022 International Symposium on Electronics and Smart Devices (ISESD)*, 2022, pp. 1–6.

[6] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition*, 2014.

[7] N. L. Baccheschi, S. Brown, J. Kerekes, and J. Schott, "Generation of a combined dataset of simulated radar and EO/IR imagery," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI*, S. S. Shen and P. E. Lewis, Eds., vol. 5806, International Society for Optics and Photonics. SPIE, 2005, pp. 88 – 99. [Online]. Available: https://doi.org/10.1117/12.605711

[8] Y. Bar-Shalom, P. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*. YBS Publishing, 2011. [Online]. Available: https://books.google.com/books?id=2aOiuAAACAAJ

[9] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[10] M. Barekatain, M. Martí, H. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2153–2160.

[11] D. H. Barker, D. T. Hodges, and T. S. Hartwick, "Far Infrared Imagery," in *Long-Wavelength Infrared*, W. L. Wolfe, Ed., vol. 0067, International Society for Optics and Photonics. SPIE, 1975, pp. 27 – 36. [Online]. Available: https://doi.org/10.1117/12.954526

[12] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.

[14] J. Becker and A. Simon, "Sensor and navigation data fusion for an autonomous vehicle," in *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511)*, 2000, pp. 156–161.

[15] L. Becker, "A review of advances in EO/IR focal plane array technology for space system applications," in *Infrared and Photoelectronic Imagers and Detector Devices II*, R. E. Longshore and A. Sood, Eds., vol. 6294, International Society for Optics and Photonics. SPIE, 2006, p. 62940R. [Online]. Available: https://doi.org/10.1117/12.684640

[16] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "YOLO-Z: improving small object detection in yolov5 for autonomous vehicles," *CoRR*, vol. abs/2112.11798, 2021. [Online]. Available: https://arxiv.org/abs/2112.11798

[17] R. Bergevin, P. St-Charles, and G. Bilodeau, "Mutual foreground segmentation with multispectral stereo pairs," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 375–384.

[18] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016.

[19] E. Bochinski, V. Eiselein, and T. Sikora, "Training a convolutional neural network for multi-class object detection using solely virtual world data," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Colorado Springs, CO, USA, Aug. 2016, pp. 278–285, electronic ISBN: 978-1-5090-3811-4 Print on Demand(PoD) ISBN: 978-1-5090-3812-1 DOI: 10.1109/AVSS.2016.7738056.

[20] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[22] T. Borvornvitchotikarn and W. Kurutach, "A taxonomy of mutual information in medical image registration," in *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2016, pp. 1–4.

[23] W. Bouachir and G.-A. Bilodeau, "Collaborative part-based tracking using salient local predictors," *Comput. Vis. Image Underst.*, vol. 137, no. C, p. 88–101, Aug. 2015. [Online]. Available: https://doi.org/10.1016/j.cviu.2015.03.010

[24] R. Breiter, J. Wendler, H. Lutz, S. Rutzinger, K. Hofmann, and J. Ziegler, "IR-detection modules from SWIR to VLWIR: performance and applications," in *Infrared Technology and Applications XXXV*, B. F. Andresen, G. F. Fulop, and P. R. Norton, Eds., vol. 7298, International Society for Optics and Photonics. SPIE, 2009, p. 72981W. [Online]. Available: https://doi.org/10.1117/12.818782

[25] J. C. Campbell, S. Wang, X. Zheng, X. Li, N. Li, F. Ma, X. Sun, C. J. Collins, A. L. Beck, B. Yang, J. B. Hurst, R. Sidhu, A. L. H.

Jr., U. Chowdhury, M. M. Wong, R. D. Dupuis, A. Huntington, L. A. Coldren, Z. Chen, E.-T. Kim, and A. Madhukar, "Photodetectors: UV to IR," in *Active and Passive Optical Components for WDM Communications III*, A. K. Dutta, A. A. S. Awwal, N. K. Dutta, and K. Fujiura, Eds., vol. 5246, International Society for Optics and Photonics. SPIE, 2003, pp. 375 – 388. [Online]. Available: https://doi.org/10.1117/12.511201

[26] R. W. Carlson, "Spectral Mapping Of Jupiter And The Galilean Satellites In The Near Infrared," in *Imaging Spectroscopy I*, D. D. Norris, Ed., vol. 0268, International Society for Optics and Photonics. SPIE, 1981, pp. 29 – 34. [Online]. Available: https://doi.org/10.1117/12.959922

[27] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220307153

[28] L. A. Chan, S. Z. Der, and N. M. Nasrabadi, "Application of dualband infrared imagery in automatic target detection," in *Automatic Target Recognition X*, F. A. Sadjadi, Ed., vol. 4050, International Society for Optics and Photonics. SPIE, 2000, pp. 282 – 293. [Online]. Available: https://doi.org/10.1117/12.395575

[29] M. Chaverot, M. Carré, M. Jourlin, A. Bensrhair, and R. Grisel, "Object detection on thermal images: Performance of yolov4 trained on small datasets," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021*, 01 2021, pp. 207–212.

[30] C. Chen, R. Xia, Y. Liu, and Y. Liu, "A simplified dual-weighted three-layer window local contrast method for infrared small target detection," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2023.

[31] G. Chen, P. St-Charles, W. Bouachir, G. Bilodeau, and R. Bergevin, "Reproducible evaluation of pan-tilt-zoom tracking," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 2055–2059.

[32] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 936–953, 2022.

[33] N. Chen, P. Guo, S. Yan, D. Piao, and Q. Zhu, "Ultrasound-assisted NIR imaging for breast cancer detection," in *Optical Tomography and Spectroscopy of Tissue IV*, B. Chance, R. R. Alfano, B. J. Tromberg, M. Tamura, and E. M. Sevick-Muraca, Eds., vol. 4250, International Society for Optics and Photonics. SPIE, 2001, pp. 546 – 557. [Online]. Available: https://doi.org/10.1117/12.434530

[34] E. Cho, B. K. McQuiston, W. Lim, S. B. Rafol, C. Hanson, R. Nguyen, and A. Hutchinson, "Development of a visible-NIR/LWIR QWIP sensor," in *Infrared Technology and Applications XXIX*, B. F. Andresen and G. F. Fulop, Eds., vol. 5074, International Society for Optics and Photonics. SPIE, 2003, pp. 735 – 744. [Online]. Available: https://doi.org/10.1117/12.497512

[35] P. Chu, "Optimal projection for multidimensional signal detection," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 2797–2800 vol.5.

[36] M. Cicconet, M. Gutwein, K. Gunsalus, and D. Geiger, "Label free cell-tracking and division detection based on 2d time-lapse images for lineage analysis of early embryo development," *Computers in Biology and Medicine*, vol. 51, 08 2014.

[37] N. Cohen, G. Sarusi, G. Mizrachi, A. Shappir, and A. Sa'ar, "Bias-controlled NIR/LWIR QWIP-based structure for night vision and see spot," in *Infrared Technology and Applications XXIX*, B. F. Andresen and G. F. Fulop, Eds., vol. 5074, International Society for Optics and Photonics. SPIE, 2003, pp. 708 – 714. [Online]. Available: https://doi.org/10.1117/12.498641

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[39] T. Damarla and A. Mehmood, "Detection of targets using distributed multi-modal sensors with correlated observations," in *SENSORS, 2013 IEEE*, 2013, pp. 1–4.

[40] A. Dammann, S. Sand, and R. Raulefs, "Signals of opportunity in mobile radio positioning," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 549–553.

[41] P. David, "Multiple-sensor cueing using a heuristic search," in *Applications of Artificial Intelligence IX*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, M. M. Trivedi, Ed., vol. 1468, Mar. 1991, pp. 1000–1009.

[42] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "CVPR19 tracking and detection challenge: How crowded can it get?" *arXiv:1906.04567 [cs]*, Jun. 2019, arXiv: 1906.04567. [Online]. Available: http://arxiv.org/abs/1906.04567

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[44] T. Dieterle, F. Particke, L. Patino-Studencki, and J. Thielecke, "Sensor data fusion of lidar with stereo rgb-d camera for object tracking," in *2017 IEEE SENSORS*, 2017, pp. 1–3.

[45] "DRS RADA TECHNOLOIES mhr (radar)," https://www.drsrada.com/products/mhr, DRS Rada Technologies, accessed: 2023-11-8.

[46] "ECHODYNE echosheild (radar)," https://www.echodyne.com/radar-solutions/echoshield/, EchoDyne Systems, accessed: 2023-11-8.

[47] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE Press, June 2008.

[48] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[49] P.-C. Fan, W.-G. Zhang, P. Kong, Q. Gao, M.-Q. Wang, and Q.-L. Dong, "Infrared dim and small target detection based on sample data," in *2021 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2021, pp. 532–536.

[50] J. M. Fialkowski and R. C. Gauss, "A physical statistical sonar clutter model for spatially-dispersed scatterers," in *OCEANS 2015 - MTS/IEEE Washington*, 2015, pp. 1–5.

[51] J. Gavan, "Ladar/radar dual mode operation system for enhancing tracking range and accuracy," in *16th International Conference on Infrared and Millimeter Waves*, M. Q. Tran, Ed., vol. 1576, International Society for Optics and Photonics. SPIE, 1991, p. 157664. [Online]. Available: https://doi.org/10.1117/12.2297933

[52] J. D. Gibson and A. Bovik, *Handbook of Image and Video Processing*, 1st ed. USA: Academic Press, Inc., 2000.

[53] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014.

[54] A. Goldberg, T. Fisher, S. Kennerly, S. Der, and A. Chan, *Analysis of Dual-Band Infrared Imagery from the Multidomain Smart Sensor Field Test*. Army Research Lab Technical Reports, 2002.

[55] A. C. Goldberg, T. Fischer, and Z. I. Derzko, "Application of dual-band infrared focal plane arrays to tactical and strategic military problems," in *Infrared Technology and Applications XXVIII*, B. F. Andresen, G. F. Fulop, and M. Strojnik, Eds., vol. 4820, International Society for Optics and Photonics. SPIE, 2003, pp. 500 – 514. [Online]. Available: https://doi.org/10.1117/12.451014

[56] ——, "Application of dual-band infrared focal plane arrays to tactical and strategic military problems," in *Proceedings of SPIE*, vol. 4820. SPIE, 2003, pp. 500–514.

[57] A. C. Goldberg, T. Fischer, Z. I. Derzko, P. N. Uppal, and M. L. Winn, "Development of a dual-band LWIR/LWIR QWIP focal plane array for detection of buried land mines," in *Infrared Detectors and Focal Plane Arrays VII*, E. L. Dereniak and R. E. Sampson, Eds., vol. 4721, International Society for Optics and Photonics. SPIE, 2002, pp. 184 – 195. [Online]. Available: https://doi.org/10.1117/12.478846

[58] M. Gong, D. Wang, X. Zhao, H. Guo, D. Luo, and M. Song, "A review of non-maximum suppression algorithms for deep learning target detection," in *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, J. Su, J. Chu, Q. Yu, and H. Jiang, Eds., vol. 11763, International Society for Optics and Photonics. SPIE, 2021, p. 1176332. [Online]. Available: https://doi.org/10.1117/12.2586477

[59] A. A. Goshtasby, *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. USA: Wiley-Interscience, 2005.

[60] H. Hallil, C. Dejous, S. Hage-Ali, O. Elmazria, J. Rossignol, D. Stuerga, A. Talbi, A. Mazzamurro, P.-Y. Joubert, and E. Lefeuvre, "Passive resonant sensors: Trends and future prospects," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12 618–12 632, 2021.

[61] H. Hariharan, A. Koschan, B. Abidi, A. Gribok, and M. Abidi, "Fusion of visible and infrared images using empirical mode decomposition to improve face recognition," in *2006 International Conference on Image Processing*, 2006, pp. 2049–2052.

[62] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[63] R. V. L. Hartley, "Transmission of information," *The Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928.

[64] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 1106–1112.

[65] L. G. Hipwood, N. Shorrocks, C. Maxey, D. Atkinson, and N. Bezawada, "SWIR and NIR MCT arrays grown by MOVPE for astronomy applications," in *Infrared Technology and Applications XXXVIII*, B. F. Andresen, G. F. Fulop, and P. R. Norton, Eds., vol. 8353, International Society for Optics and Photonics. SPIE, 2012, p. 83532M. [Online]. Available: https://doi.org/10.1117/12.919007

[66] J. Hird, "Multiresolution object detection and segmentation using top-down algorithms," in *Third International Conference on Image Processing and its Applications, 1989.*, 1989, pp. 416–420.

[67] S. M. Hong, H. Lee, I. Baek, and M. S. Kim, "MCT-based SWIR hyperspectral imaging system for evaluation of biological samples ," in *Sensing for Agriculture and Food Quality and Safety VIII*, M. S. Kim, K. Chao, and B. A. Chin, Eds., vol. 9864, International Society for Optics and Photonics. SPIE, 2016, p. 986410. [Online]. Available: https://doi.org/10.1117/12.2227177

[68] Y. Huang, Z. Jiang, R. Lan, S. Zhang, and K. Pi, "Infrared image super-resolution via transfer learning and psrgan," *IEEE Signal Processing Letters*, vol. 28, pp. 982–986, 2021.

[69] R. Hudson and J. Hudson, "The military applications of remote sensing by infrared," *Proceedings of the IEEE*, vol. 63, no. 1, pp. 104–128, 1975.

[70] B. L. Huneycutt, "Shuttle Imaging Radar - B/C instruments," in *Instrumentation for Optical Remote Sensing from Space*, J. W. Lear, A. Monfils, S. L. Russak, and J. S. Seeley, Eds., vol. 0589, International Society for Optics and Photonics. SPIE, 1986, pp. 137 – 157. [Online]. Available: https://doi.org/10.1117/12.951925

[71] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[72] Hwann-Tzong Chen, Horng-Horng Lin, and Tyng-Luh Liu, "Multi-object tracking using dynamical graph matching," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2, Dec 2001, pp. II–II.

[73] T. Imaging, "Flir data set dataset," https://universe.roboflow.com/thermal-imaging-0hwfw/flir-data-set , jun 2023, visited on 2023-06-04. [Online]. Available: https://universe.roboflow.com/thermal-imaging-0hwfw/flir-data-set

[74] J. A. Jamieson, "Special electronic circuits for nonimage-forming infrared systems," *Proceedings of the IRE*, vol. 47, no. 9, pp. 1570–1572, 1959.

[75] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[76] M. Jiang, P. P. Shum, B. Lin, S. C. Tjin, and Y. Jiang, "A stable dual-wavelength single-longitudinal-mode fiber laser with a tunable wavelength spacing based on a chirped phase-shifted grating filter," in *Passive Components and Fiber-Based Devices VIII*, B. P. Pal, Ed., vol. 8307, International Society for Optics and Photonics. SPIE, 2011, p. 83070M. [Online]. Available: https://doi.org/10.1117/12.904451

[77] Z. Jie, C. Dong, and S. Dewei, "K distribution sea clutter modeling and simulation based on zmnl," in *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2015, pp. 506–509.

[78] G. Jocher, "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," Oct. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4154370

[79] V. J. N. Jr., R. S. Evans, and D. G. Currie, "Performance comparison of visual, infrared, and ultraviolet sensors for landing aircraft in fog," in *Enhanced and Synthetic Vision 1999*, J. G. Verly, Ed., vol. 3691, International Society for Optics and Photonics. SPIE, 1999, pp. 2 – 20. [Online]. Available: https://doi.org/10.1117/12.354430

[80] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, March 2004.

[81] M. S. Khan, G. A. Washer, and S. B. Chase, "Evaluation of dual-band infrared thermography system for bridge deck delamination surveys," in *Structural Materials Technology III: An NDT Conference*, R. D. Medlock and D. C. Laffrey, Eds., vol. 3400, International Society for Optics and Photonics. SPIE, 1998, pp. 224 – 235. [Online]. Available: https://doi.org/10.1117/12.300094

[82] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with language referring expressions," in *ACCV*, 2018.

[83] S.-H. Kim and H.-G. Kim, "Face detection using multi-modal information," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 14–19.

[84] S. Kim, H. Kim, W. Yoo, and K. Huh, "Sensor fusion algorithm design in detecting vehicles using laser scanner and stereo vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1072–1084, 2016.

[85] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," *CoRR*, vol. abs/1902.07296, 2019. [Online]. Available: http://arxiv.org/abs/1902.07296

[86] P. Konstantinova, A. Udvarev, and T. Semerdjiev, "A study of a target tracking algorithm using global nearest neighbor approach," *International Conf. on Computer Systems and Technologies*, 01 2003.

[87] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[88] J. Kreer, "A question of terminology," *IRE Transactions on Information Theory*, vol. 3, no. 3, pp. 208–208, 1957.

[89] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, A. Gupta, A. Bibi, A. Lukezic, A. Garcia-Martin, A. Saffari, A. Petrosino, and A. Solis Montero, "The visual object tracking vot2015 challenge results," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 564–586.

[90] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojíř, G. Fernández, A. Lukežič, A. Dimitriev, A. Petrosino, A. Saffari, B. Li, B. Han, C. Heng, C. Garcia, D. Pangeršič, G. Häger, F. S. Khan, F. Oven, H. Possegger, H. Bischof, H. Nam, J. Zhu, J. Li, J. Y. Choi, J.-W. Choi, J. F. Henriques, J. van de Weijer, J. Batista, K. Lebeda, K. Öfjäll, K. M. Yi, L. Qin, L. Wen, M. E. Maresca, M. Danelljan, M. Felsberg, M.-M. Cheng, P. Torr, Q. Huang, R. Bowden, S. Hare, S. Y. Lim, S. Hong, S. Liao, S. Hadfield, S. Z. Li, S. Duffner, S. Golodetz, T. Mauthner, V. Vineet, W. Lin, Y. Li, Y. Qi, Z. Lei, and Z. H. Niu, "The visual object tracking vot2014 challenge results," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015, pp. 191–217.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[92] M. Krišto, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using yolo," *IEEE Access*, vol. 8, pp. 125 459–125 476, 2020.

[93] P. LACOMME, J.-P. HARDANGE, J.-C. MARCHAIS, and E. NORMANT, "19 - radar applications and roles," in *Air and Spaceborne Radar Systems*, P. LACOMME, J.-P. HARDANGE, J.-C. MARCHAIS, and E. NORMANT, Eds. Norwich, NY: William Andrew Publishing, 2001, pp. 347–369.

[94] M. Laurenzis, F. Christnacher, and A. Velten, "Study of a dual mode SWIR active imaging system for direct imaging and non-line-of-sight vision," in *Laser Radar Technology and Applications XX; and Atmospheric Propagation XII*, M. D. Turner, G. W. Kamerman, L. M. W. Thomas, and E. J. Spillar, Eds., vol. 9465, International Society for Optics and Photonics. SPIE, 2015, p. 946509. [Online]. Available: https://doi.org/10.1117/12.2175857

[95] R. Layton and D. Witkowski, "5G Versus Wi-Fi: Challenges for Economic, Spectrum, and Security Policy," *Journal of Information Policy*, vol. 11, pp. 523–561, 12 2021. [Online]. Available: https://doi.org/10.5325/jinfopoli.11.2021.0523

[96] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

[97] H. D. Lee, C. S. Kim, M. Y. Jeong, and Z. Chen, "Dual-band wavelength-swept active mode locking laser for multi-band fiber-optic sensors," in *OFS2012 22nd International Conference on Optical Fiber Sensors*, Y. Liao, W. Jin, D. D. Sampson, R. Yamauchi, Y. Chung, K. Nakamura, and Y. Rao, Eds., vol. 8421, International Society for Optics and Photonics. SPIE, 2012, p. 84215S. [Online]. Available: https://doi.org/10.1117/12.975123

[98] J. Lee and C. Lin, "A novel approach to real-time motion detection," in *Proceedings CVPR '88: The Computer Society Conference on Computer Vision and Pattern Recognition*, 1988, pp. 730–735.

[99] C. Leigh and A. Richardson, "Performance analysis of an infrared pilot-warning indicator system," *Proceedings of the IEEE*, vol. 58, no. 3, pp. 462–469, 1970.

[100] C. L. Leonard, M. J. DeWeert, J. Gradie, J. Iokepa, and C. L. Stalder, "Performance of an EO/IR sensor system in marine search and rescue," in *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications II*, S. H. Wyatt, Ed., vol. 5787, International Society for Optics and Photonics. SPIE, 2005, pp. 122 – 133. [Online]. Available: https://doi.org/10.1117/12.603909

[101] E. Letsoalo and S. Ojo, "A model to mitigate session hijacking attacks in wireless networks," in *2018 IST-Africa Week Conference (IST-Africa)*, 2018, pp. Page 1 of 10–Page 10 of 10.

[102] J. M. Lewis, S. Lakshmivarahan, and S. Dhall, *Dynamic Data Assimilation: A Least Squares Approach*, ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006.

[103] A. Li, M. Lin, Y. Wu, M. Yang, and S. Yan, "NUS-PRO: A New Visual Tracking Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 335–349, 2016.

[104] A. Li, Z. Chen, and Y. Wang, "Buaa-pro: A tracking dataset with pixel-level annotation," in *British Machine Vision Conference*, 2018.

[105] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "Yolov6: A single-stage object detection framework for industrial applications," 2022.

[106] S. Li, Y. Li, Y. Li, M. Li, and X. Xu, "Yolo-firi: Improved yolov5 for infrared image object detection," *IEEE Access*, vol. 9, pp. 141 861–141 875, 2021.

[107] C. Liang, Z. Zhang, Y. Lu, X. Zhou, B. Li, X. Ye, and J. Zou, "Rethinking the competition between detection and reid in multi-object tracking," *CoRR*, vol. abs/2010.12138, 2020. [Online]. Available: https://arxiv.org/abs/2010.12138

[108] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5630–5644, Dec 2015.

[109] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[110] D. Liu, H. Mansour, and P. T. Boufounos, "Robust mutual information-based multi-image registration," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 915–918.

[111] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016.

[112] Q. Liu, Z. Li, and S. Li, "Performance analysis on vibe detection of moving object," in *2022 International Conference on Automation, Robotics and Computer Engineering (ICARCE)*, 2022, pp. 1–5.

[113] S. Liu, Y. Yang, Q. Li, H. Feng, Z. Xu, Y. Chen, and L. Liu, "Infrared image super resolution using gan with infrared image prior," in *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, 2019, pp. 1004–1009.

[114] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[115] M. I. A. Lourakis and A. A. Argyros, "Sba: A software package for generic sparse bundle adjustment," *ACM Trans. Math. Softw.*, vol. 36, pp. 2:1–2:30, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:474253

[116] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[117] H. Luo, P. Wang, H. Chen, and V. P. Kowelo, "Small object detection network based on feature information enhancement," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[118] L. Ma, J. Wu, J. Zhang, Z. Wu, G. Jeon, Y. Zhang, and T. Wu, "Research on sea clutter reflectivity using deep learning model in industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 5929–5937, 2020.

[119] T. Ma, Z. Yang, J. Wang, S. Sun, X. Ren, and U. Ahmad, "Infrared small target detection network with generate label and feature mapping," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[120] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, 2nd ed. North-holland Publishing Company, 1978.

[121] U. Madhow, *Introduction to Communication Systems*. Cambridge University Press, 2014. [Online]. Available: https://books.google.com/books?id=XoBIBQAAQBAJ

[122] 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, 1997.

[123] V. Maik, H. Kim, D. Kim, E. Chae, and J. Paik, "Robust background generation using a modified mixture of gaussian model for object detection," in *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, June 2014, pp. 1–2.

[124] A. Manzanera, "σ-δ background subtraction and the zipf law," in *Progress in Pattern Recognition, Image Analysis and Applications*, L. Rueda, D. Mery, and J. Kittler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 42–51.

[125] Y. Mao, S. Chang, E. Murdock, and C. Flueraru, "Simultaneous 1310/1550 dual-band swept laser source and fiber-based dual-band common-path swept source optical coherence tomography," in *Photonics North 2011*, R. Kashyap, M. Têtu, and R. N. Kleiman, Eds., vol. 8007, International Society for Optics and Photonics. SPIE, 2011, p. 800704. [Online]. Available: https://doi.org/10.1117/12.902617

[126] D. A. Masliah, B. S. Cole, A. Platzker, and M. Schindler, "High-efficiency dual-band power amplifier for radar applications," in *Monolithic Microwave Integrated Circuits for Sensors, Radar, and Communications Systems*, R. F. Leonard and K. B. Bhasin, Eds., vol. 1475, International Society for Optics and Photonics. SPIE, 1991, pp. 113 – 120. [Online]. Available: https://doi.org/10.1117/12.44486

[127] R. McMillan, "Surveillance technologies," in *EASCON '88., 21st Annual Electronics and Aerospace Conference, How will Space and Terrestrial Systems Share the Future?*, 1988, pp. 55–63.

[128] H. Medeiros, J. Park, and A. Kak, "Distributed object tracking using a cluster-based kalman filter in wireless camera networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 448–463, 2008.

[129] A. D. Meigs, L. J. O. III, T. Y. Cherezova, B. Rafert, and R. G. Sellar, "LWIR and MWIR ultraspectral Fourier transform imager," in *Sensors, Systems, and Next-Generation Satellites*, H. Fujisada, Ed., vol. 3221, International Society for Optics and Photonics. SPIE, 1997, pp. 421 – 428. [Online]. Available: https://doi.org/10.1117/12.298109

[130] O. Mendoza-Schrock, J. A. Patrick, and E. P. Blasch, "Video image registration evaluation for a layered sensing environment," in *Proceedings of the IEEE 2009 National Aerospace & Electronics Conference (NAECON)*, 2009, pp. 223–230.

[131] O. Mendoza-Schrock, J. A. Patrick, and M. Garing, "Exploring image registration techniques for layered sensing," in *Evolutionary and Bio-Inspired Computation: Theory and Applications III*, T. H. O'Donnell, M. Blowers, and K. L. Priddy, Eds., vol. 7347, International Society for Optics and Photonics. SPIE, 2009, pp. 292 – 306. [Online]. Available: https://doi.org/10.1117/12.818526

[132] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[133] L. G. Minor and J. Sklansky, "The detection and segmentation of blobs in infrared images," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 3, pp. 194–201, 1981.

[134] V. V. Molebny, I. G. Pallikaris, L. P. Naoumidis, G. W. Kamerman, E. M. Smirnov, L. M. Ilchenko, and V. O. Goncharov, "Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles," in *Laser Radar Technology and Applications*, G. W. Kamerman, Ed., vol. 2748, International Society for Optics and Photonics. SPIE, 1996, pp. 68 – 75. [Online]. Available: https://doi.org/10.1117/12.243574

[135] ——, "Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles," in *Laser Radar Technology and Applications*, G. W. Kamerman, Ed., vol. 2748, International Society for Optics and Photonics. SPIE, 1996, pp. 68 – 75. [Online]. Available: https://doi.org/10.1117/12.243574

[136] T. P. Morton and P. A. Maresca, "Small target detection by space-borne radiometry," in *1983 Eighth International Conference on Infrared and Millimeter Waves*, 1983, pp. 1–2.

[137] J. Mou, W. Gao, and Z. Song, "Image fusion based on non-negative matrix factorization and infrared feature extraction," in *2013 6th International Congress on Image and Signal Processing (CISP)*, vol. 2, 2013, pp. 1046–1050.

[138] A. Moudgil and V. Gandhi, "Long-term visual object tracking benchmark," *CoRR*, vol. abs/1712.01358, 2017. [Online]. Available: http://arxiv.org/abs/1712.01358

[139] C. T. Nguyen, J. P. Havlicek, G. Fan, J. T. Caulfield, and M. S. Pattichis, "Robust dual-band mwir/lwir infrared target tracking," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 78–83.

[140] D.-L. Nguyen, P.-L. St-Charles, and G.-A. Bilodeau, "Non-planar infrared-visible registration for uncalibrated stereo pairs," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 329–337.

[141] L. Nolibe, J. Borgnino, M. Ducoulombier, and M. Artaud, "Adaptive multispectral detection of small targets using spatial and spectral convergence factor," in *Signal and Data Processing of Small Targets 1996*, O. E. Drummond, Ed., vol. 2759, International Society for

Optics and Photonics. SPIE, 1996, pp. 111 – 120. [Online]. Available: https://doi.org/10.1117/12.241161

[142] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187 – 1200, Jun 2014, preprint. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2014/OB14b

[143] M. E. O'Hanlon, "Forecasting change in military technology, 2020-2040," *Military Technology*, vol. 2020, p. 2040, 2018.

[144] J.-W. Perng, P.-Y. Liu, K.-Q. Zhong, and Y.-W. Hsu, "Front object recognition system for vehicles based on sensor fusion using stereo vision and laser range finder," in *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, 2017, pp. 261–262.

[145] W. Peterson, T. Birdsall, and W. Fox, "The theory of signal detectability," *Transactions of the IRE Professional Group on Information Theory*, vol. 4, no. 4, pp. 171–212, 1954.

[146] V. Petrushevsky, Y. Karklinsky, and A. Chernobrov, "ElOp EO/IR LOROP camera: image stabilization for dual-band whiskbroom scanning photography," in *Infrared Technology and Applications XXVIII*, B. F. Andresen, G. F. Fulop, and M. Strojnik, Eds., vol. 4820, International Society for Optics and Photonics. SPIE, 2003, pp. 607 – 617. [Online]. Available: https://doi.org/10.1117/12.453838

[147] "PVP ADVANCED EO SYSTEMS night hawk positioner (gimbals)," https://advancedeo.systems/products/category.aspx?categoryID=2, PVP Advanced EO Systems, accessed: 2023-11-6.

[148] S. Quan, W. Qian, J. Guo, and H. Zhao, "Visible and infrared image fusion based on curvelet transform," in *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*, 2014, pp. 828–832.

[149] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.

[150] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[151] ——, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[152] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[153] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[154] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[155] D. Ruck, S. Rogers, M. Kabrisky, and J. Mills, "Target recognition: Conventional and neural network approaches," in *International 1989 Joint Conference on Neural Networks*, 1989, pp. 608 vol.2–.

[156] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[157] A. Sanna, B. Pralio, F. Lamberti, and G. Paravati, "A novel ego-motion compensation strategy for automatic target tracking in flir video sequences taken from uavs," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 2, pp. 723–734, 2009.

[158] R. G. Sementelli, "EO/IR dual-band reconnaissance system DB-110," in *Airborne Reconnaissance XIX*, W. G. Fishell, A. A. Andraitis, P. A. Henkel, and A. C. C. Jr., Eds., vol. 2555, International Society for Optics and Photonics. SPIE, 1995, pp. 222 – 231. [Online]. Available: https://doi.org/10.1117/12.218615

[159] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," 2014.

[160] P. Shaniya, G. Jati, M. R. Alhamidi, W. Caesarendra, and W. Jatmiko, "Yolov4 rgbt human detection on unmanned aerial vehicle perspective,"

in *2021 6th International Workshop on Big Data and Information Security (IWBIS)*, 2021, pp. 41–46.

[161] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[162] M. Shao, Y. Wang, and Y. Wang, "A super-resolution based method to synthesize visual images from near infrared," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2453–2456.

[163] P. Sharma, K. K. Sarma, and N. E. Mastorakis, "Artificial intelligence aided electronic warfare systems- recent trends and evolving applications," *IEEE Access*, vol. 8, pp. 224 761–224 780, 2020.

[164] L. Shu and T. Tan, "Sar and spot image registration based on mutual information with contrast measure," in *2007 IEEE International Conference on Image Processing*, vol. 5, 2007, pp. V – 429–V – 432.

[165] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, July 2014.

[166] M. Sohail, A. Gilgiti, and T. Rahman, "Ultrasonic and stereo vision data fusion," in *8th International Multitopic Conference, 2004. Proceedings of INMIC 2004.*, 2004, pp. 357–361.

[167] S. Sonn, G.-A. Bilodeau, and P. Galinier, "Fast and accurate registration of visible and infrared videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 308–313.

[168] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[169] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking." in *CVPR*. IEEE Computer Society, 1999, pp. 2246–2252.

[170] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," in *Sensor Fusion: Architectures, Algorithms, and Applications III*, B. V. Dasarathy, Ed., vol. 3719, International Society for Optics and Photonics. SPIE, 1999, pp. 430 – 441. [Online]. Available: https://doi.org/10.1117/12.341367

[171] S. Stergiopoulos, *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*, ser. CRC Press Revivals. CRC Press, 2017. [Online]. Available: https://books.google.com/books?id=mXFQDwAAQBAJ

[172] B. Stuckman, G. Zimmerman, and C. Perttunen, "A solid state infrared device for detecting the presence of car in a driver's blind spot," in *Proceedings of the 32nd Midwest Symposium on Circuits and Systems,*, 1989, pp. 1185–1188 vol.2.

[173] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, 2003.

[174] T.Brox and J.Malik, "Object segmentation by long term analysis of point trajectories," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science. Springer, Sept. 2010. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2010/Bro10c

[175] "Free teledyne flir thermal dataset for algorithm training," https://https://www.flir.com/oem/adas/adas-dataset-form/#anchor29, Teledyne FLIR Systems, accessed: 2022-10-30.

[176] "TELEDYNE FLIR flir ptu-d48e," https://www.flir.com/products/ptu-d48e/?vertical=mcs&segment=oem, Teledyne FLIR Systems, accessed: 2023-11-6.

[177] G. A. Tidhar, O. B. Aphek, and E. Cohen, "OTHELLO: a novel SWIR dual-band detection system and its applications," in *Infrared Technology and Applications XXXIX*, B. F. Andresen, G. F. Fulop, C. M. Hanson, P. R. Norton, and P. Robert, Eds., vol. 8704, International Society for Optics and Photonics. SPIE, 2013, p. 87040E. [Online]. Available: https://doi.org/10.1117/12.2019202

[178] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image and Vision Computing*, vol. 97, p. 103910, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885620300421

[179] S. Toral, M. Vargas, F. Barrero, and M. G. Ortega, "Improved sigma-delta background estimation for vehicle detection," *Electronics Letters*, vol. 45, no. 1, pp. 32–34, January 2009.

[180] T. Toriu and H. Fukumoto, "A learning method for association between vision and ego-motion which is capable of adapting to arbitrary image distortion," in *2008 23rd International Conference Image and Vision Computing New Zealand*, 2008, pp. 1–6.

[181] J. R. Tower, B. M. McCarthy, L. E. PelIon, R. T. Strong, H. Elabd, A. D. Cope, D. M. Hoffman, W. M. Kramer, and R. W. Longsderff, "Visible And Shortwave Infrared Focal Planes For Remote Sensing Instruments," in *Recent Advances in Civil Space Remote Sensing*, vol. 0481, International Society for Optics and Photonics. SPIE, 1984, pp. 24 – 33. [Online]. Available: https://doi.org/10.1117/12.943065

[182] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, 2013. [Online]. Available: http://www.huppelen.nl/publications/selectiveSearchDraft.pdf

[183] F. Vachss, J. B. Norman, and C. E. T. Jr., "Dual band mid-wave/long-wave IR source for atmospheric remote sensing," in *Air Monitoring and Detection of Chemical and Biological Agents*, J. Leonelli and M. L. Althouse, Eds., vol. 3533, International Society for Optics and Photonics. SPIE, 1999, pp. 174 – 179. [Online]. Available: https://doi.org/10.1117/12.336855

[184] H. Visser and B. A. van der Zwan, "Design and development of a dual-wavelength satellite laser ranging system," in *Design and Engineering of Optical Systems*, J. J. M. Braat, Ed., vol. 2774, International Society for Optics and Photonics. SPIE, 1996, pp. 655 – 666. [Online]. Available: https://doi.org/10.1117/12.246709

[185] J. Vivekanandan, F. J. Turk, and V. N. Bringi, "Remote sensing of precipitation structures using combined microwave radar and radiometric techniques," in *Wave Propagation and Scattering in Varied Media II*, V. K. Varadan, Ed., vol. 1558, International Society for Optics and Photonics. SPIE, 1991, pp. 324 – 338. [Online]. Available: https://doi.org/10.1117/12.49638

[186] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.

[187] K. Wang, N. Ding, and F. Dai, "A simple and parallel algorithm for robot position estimation by stereo visual-inertial sensor fusion," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2017, pp. 860–864.

[188] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," *CoRR*, vol. abs/1909.12605, 2019. [Online]. Available: http://arxiv.org/abs/1909.12605

[189] A. R. Weiß, U. Adomeit, P. Chevalier, S. Landeau, P. Bijl, F. Champagnat, J. Dijk, B. Göhler, S. Landini, J. P. Reynolds, and L. N. Smith, "A standard data set for performance analysis of advanced IR image processing techniques," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXIII*, G. C. Holst and K. A. Krapels, Eds., vol. 8355, International Society for Optics and Photonics. SPIE, 2012, p. 835512. [Online]. Available: https://doi.org/10.1117/12.919004

[190] A. R. Weiß, U. Adomeit, P. Chevalier, S. Landeau, P. Bijl, F. Champagnat, J. Dijk, B. Göhler, S. Landini, J. P. Reynolds, and L. N. Smith, "A standard data set for performance analysis of advanced IR image processing techniques," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXIII*, G. C. Holst and K. A. Krapels, Eds., vol. 8355, International Society for Optics and Photonics. SPIE, 2012, pp. 354 – 363. [Online]. Available: https://doi.org/10.1117/12.919004

[191] K. L. Wong, "GITHUB yolov78 (wongkinliu)," https://https://github.com/WongKinYiu/yolov7, accessed: 2023-11-8.

[192] J. Woo, J.-H. Baek, S.-H. Jo, S. Y. Kim, and J.-H. Jeong, "A study on object detection performance of yolov4 for autonomous driving of tram," *Sensors*, vol. 22, no. 22, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/22/9026

[193] M. Wu, X. Yang, Z. Fu, H. He, J. Du, T. Xu, and Z. Tu, "Infrared moving small target detection based on consistency of sparse trajectory," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–1, 2023.

[194] R. Wu, D. Yu, J. Liu, H. Wu, W. Chen, and Q. Gu, "An improved fusion method for infrared and low-light level visible image," in *2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2017, pp. 147–151.

[195] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark." in *CVPR*. IEEE Computer Society, 2013, pp. 2411–2418.

[196] ——, "Object tracking benchmark." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, 2015.

[197] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu, "Deep learning for image inpainting: A survey," *Pattern Recognition*, vol. 134, p. 109046, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032200526X

[198] F. Xiangsuo, Q. Wenlin, L. Juliu, H. Qingnan, and Z. Fan, "Dim and small target detection based on spatio-temporal filtering and high-order energy estimation," *IEEE Photonics Journal*, vol. 15, no. 2, pp. 1–20, 2023.

[199] W. Xiong, L. Xiang, J. Li, and X. Zhao, "Moving object detection algorithm based on background subtraction and frame differencing," in *Proceedings of the 30th Chinese Control Conference*, 2011, pp. 3273–3276.

[200] W. Xu-ming, C. Jin-yan, and Z. Ben, "Nonlinear modeling and simulation of radar clutter," in *2010 Second International Conference on Computer Modeling and Simulation*, vol. 3, 2010, pp. 421–424.

[201] A. Yan, J. Li, Y. Wang, Y. Xue, and X. Sun, "Research on moving target detection based on improved gaussian mixture model," in *2020 Chinese Control And Decision Conference (CCDC)*, 2020, pp. 1168–1173.

[202] W. Yong-liang, X. Wen-chong, D. Ke-qing, and Z. Xi-chuan, "General clutter modeling for airborne radar," in *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, 2010, pp. 2274–2278.

[203] S. Zhang, Z. Liu, B. Liu, and F. Zhou, "Medical image registration by using salient phase congruency and regional mutual information," in *2011 4th International Congress on Image and Signal Processing*, vol. 2, 2011, pp. 760–764.

[204] X. Zhang, C. Leng, Y. Hong, Z. Pei, I. Cheng, and A. Basu, "Multimodal remote sensing image registration methods and advancements: A survey," *Remote Sensing*, vol. 13, no. 24, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/24/5128

[205] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1838–1845.

[206] Y. Zhou, K. Gao, Z. Dou, Z. Hua, and H. Wang, "Target-aware fusion of infrared and visible images," *IEEE Access*, vol. 6, pp. 79 039–79 049, 2018.

Appendices

Appendix A

Parallel ViBe

```
% Create Samples
mw_gpuFrames = gpuArray(mw_frame_storage);
lw_gpuFrames = gpuArray(lw_frame_storage);

mw_gpuCount =
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));
lw_gpuCount =
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));
mw_detections =
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));
lw_detections =
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));
% Bootstrap bg model here....
mw_gpuSamples = gpuArray(mw_frame_storage(:,:,1:obj.N));
lw_gpuSamples = gpuArray(lw_frame_storage(:,:,1:obj.N));

p_min = obj.pound_min;

%% Define storage for ViBE
lw_true_positives =
        gpuArray(zeros(1,obj.frame_count));
lw_false_positives =
        gpuArray(zeros(1,obj.frame_count));
mw_true_positives =
        gpuArray(zeros(1,obj.frame_count));
mw_false_positives =
        gpuArray(zeros(1,obj.frame_count));
tc_true_positives =
        gpuArray(zeros(1,obj.frame_count));
tc_false_positives =
        gpuArray(zeros(1,obj.frame_count));
```

```
%% Start processing
for frame_idx=1:obj.frame_count

        % Initialize some variables....
        mw_gpuCount = ...
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));

        lw_gpuCount = ...
        gpuArray(zeros(size(mw_frame_storage(:,:,1))));
        foreground_map = ...
         gpuArray(zeros(size(mw_frame_storage(:,:,1))));

    for samp_idx = 1:N_samples
     if((abs(samp_idx - frame_idx) > 3))
          % ^^^ Helps with initializations
              mw_gpuCount = mw_gpuCount +
                   (abs(mw_gpuSamples(:,:,samp_idx) -
                   mw_gpuFrames(:,:,frame_idx)) < obj.R);

                   lw_gpuCount = lw_gpuCount + ...
                   (abs(lw_gpuSamples(:,:,samp_idx) - ...
                   lw_gpuFrames(:,:,frame_idx)) < obj.R);
        else
                   mw_gpuCount = mw_gpuCount + 1;
                   lw_gpuCount = lw_gpuCount + 1;
        end
    end

        % Generate pixel-wise detections
        mw_detections = (mw_gpuCount < p_min);
        lw_detections = (lw_gpuCount < p_min);
        [m n] = size(mw_detections);

        %% MW Learning
        learning_locations =
        (floor(8*rand(m,n,"gpuArray")) == 0) &
        ~mw_detections;

        eaten_up_location =
        (floor(16*rand(m,n,"gpuArray")) == 0) &
```

```
mw_detections;

learning_frame_number =
        mod(frame_idx-2,obj.N)+1;

tmp_sample =
  mw_gpuSamples(:,:,learning_frame_number);
tmp_frame =
  mw_gpuFrames(:,:,frame_idx);
tmp_sample(learning_locations | ...
        eaten_up_location) =
        tmp_frame(learning_locations | ...
         eaten_up_location);

mw_gpuSamples(:,:,learning_frame_number) =
 tmp_sample;

%% LW Learning
learning_locations =
 (floor(8*rand(m,n,"gpuArray")) == 0) ...
 & ~lw_detections;
eaten_up_location =
 (floor(16*rand(m,n,"gpuArray")) == 0) & ...
 lw_detections;
tmp_sample =
 lw_gpuSamples(:,:,learning_frame_number);
tmp_frame = lw_gpuFrames(:,:,frame_idx);
tmp_sample(learning_locations | ...
                eaten_up_location) = ...
  tmp_frame(learning_locations | ...
   eaten_up_location);
twocolor_detections = lw_detections | ...
                mw_detections;

% Create Two-Color Motion Detections by
% Or-ing two detections together
twocolor_detections = lw_detections | ...
                 mw_detections;

%% Morphological Cleaning
```

```
structuring_element2 =
        [0 0 1 0 0;
         0 1 1 1 0;
         1 1 1 1 1;
         0 1 1 1 0;
         0 0 1 0 0];

structuring_element =
        [ 0 1 0;
          1 1 1;
          0 1 0 ];

mw_detections_despeckled = ...
imdilate(
imerode(mw_detections,structuring_element),
structuring_element2);
lw_detections_despeckled = ...
imdilate(
imerode(lw_detections,structuring_element),
structuring_element2);

%% label connected components
[lw_label,lw_n] =
        bwlabel(lw_detections_despeckled);
[mw_label,mw_n] =
        bwlabel(mw_detections_despeckled);
[tc_label,tc_n] =
        bwlabel(mw_detections_despeckled| ...
lw_detections_despeckled);

%% lw build bounding boxes
lw_dets = zeros(lw_n,6);
% Extra columns for plottting compatibility
for blob_idx=1:lw_n
        [row, col] = find(lw_label == blob_idx);
        % move out of gpu
        row = gather(row);
        col = gather(col);
        lw_dets(blob_idx,1) = blob_idx;
        lw_dets(blob_idx,4) = min(row);
```

```
                lw_dets(blob_idx,3) = min(col);
                lw_dets(blob_idx,6) = max(row)
                        - min(row) + 1;
                lw_dets(blob_idx,5) = max(col)
                        - min(col) + 1;
        end

        % Generate a frame index for later evaluation.
        lw_dets(:,2) = frame_idx;

        ... mw and tc bounding boxes are generated
         the same way.
end
```

Please note that the version presented here lacks the implementation of neighborhood learning, as seen in the original text. The following code snippet is provided to offer the reader insight into how to reintegrate that functionality:

```
%% Neighborhood sharing
        % Generate eight neighbor location
        neighbor_learning_directions = ...
          (floor(8*rand(m,n,"gpuArray") + 1)
           & learning_locations;
        % Upper left learning
        [row,col] = ind2sub( ...
                size(neighbor_learning_directions), ...
        find(neighbor_learning_directions == 1));
        col = col(row - 1 > 0);
        row = row(row - 1 > 0);
        gpuSamples(row,col, ...
                some_appropriate_frame_number) =  ...
                gpuSamples(row-1,col,
                some_appropriate_frame_number);
```

Appendix B

SQL Script

```
select
sequence_id,
frame_number,
`object`,
det_idx as 'gt2',
intersection/(union_plus-intersection) as IoU,
(gt_w+1)*(gt_h+1) as 'object_area',
gt_idx as 'gt1'
from
(select
sequence_id,
frame_number,
`object`,
det_idx,
x_min,
x_max,
y_min,
y_max,
gt_w,
gt_h,
if(x_max-x_min+1 > 0,x_max-x_min+1, 0)
*if(y_max-y_min+1 > 0,
y_max-y_min+1, 0) as `intersection`,
cast((det_w+1)*(det_h+1) + (gt_w+1)
*(gt_h+1) as float)
as union_plus,
gt_idx
from
(select
g.`object`,
g.frame_number,
d.sequence_id,
```

```
d.pk as 'det_idx',
cast(if(d.x > g.x,d.x,g.x) as float)
as x_min,
cast(if(d.y > g.y,d.y,g.y) as float)
as y_min,
cast(if(d.x + d.width  > g.x + g.width,
g.x + g.width,
d.x + d.width) as float) as x_max,
cast(if(d.y + d.height  > g.y + g.height,
g.y + g.height,
d.y + d.height) as float) as y_max,
d.width as 'det_w',
d.height as 'det_h',
g.width as 'gt_w',
g.height as 'gt_h',
g.pk as 'gt_idx'
from two_color.gt_w_idx d join
two_color.gt_w_idx  g on
g.sequence_id  = d.sequence_id
AND
g.frame_number = d.frame_number) s) ss
where det_idx <> gt_idx
and intersection/(union_plus-intersection)
> 0
order by sequence_id, frame_number
```

Acronyms

ANN Artificial Neural Network

AP Average Precision

ARL Army Research Lab

ATR Automatic Target Recognition

CCA Connected Components Analysis

CNN Convolutional Neural Network

COA Common Optical Axis

DBIR Dual-Band Infrared

DBIR Dual-Band Infrared

DL Deep Learning

EKF Extended Kalman Filter

EO Electro-Optical

FIR Far Infrared

FoV Field-of-View

FP False Positive (FP)

FPA Focal-Plane Array

FPS frames per second

GAN Generative Adversarial Network

GMM Gaussian Mixture Model

IR Infrared

JDL Joint Directors of Laboratories

KF Kalman Filter

LWIR Long-Wave Infrared

mAP Mean Average Precision

MHT Multiple Hypothesis Tracker

MLP Multi-Layer Perceptron

MOT Multi-Object Tracker

MS COCO Microsoft Common Objects in COntext

MW Mid-Wave

MWIR Mid-Wave Infrared

NIR Near Infrared

PASCAL VOC Connected Components Analysis

PF Particle-Filter

R-CNN Region Convolution Neural Network

RS Remote Sensing

SPIE Society of Photo-Optical Instrumentation Engineers

SR Super-Resolution

SWIR Short-Wave Infrared

UKF Unscented Kalman Filter

ViBe Visual Background Extractor

VIS Visible-spectrum

YOLO You Only Look Once

YOLOv3 You Only Look Once version 3

YOLOv4 You Only Look Once version 4

YOLOv7 You Only Look Once version 7