Computer vision-based post-disaster needs assessment from low altitude aerial imagery

by

René Andrés García Franceschini

B.S. Civil and Environmental Engineering,
Massachusetts Institute of Technology (2019)

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Institute for Data, Systems, and Society
May 14, 2021

Certified by .....
Saurabh Amin
Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Noelle Eckley Selin
Associate Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric and Planetary Sciences
Director, Technology and Policy Program

Computer vision-based post-disaster needs assessment from low altitude aerial imagery

by

René Andrés García Franceschini

Submitted to the Institute for Data, Systems, and Society
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

Over the past decades, climate change has driven an increase in the frequency and intensity of natural disasters. In an effort to increase the situational awareness and timely support for search and rescue missions in the aftermath of a disaster, the United States Civil Air Patrol (CAP) gathers aerial imagery of the impacted areas. However, these high resolution and timely images are seldom used for quantitative assessment of damage. This thesis focuses on the following question: **How can we use modern computer vision techniques to utilize CAP imagery for post-disaster needs assessment, specifically for the purpose of damage estimation and localization?** This question is important because the data gathered by CAP has significant potential to expedite response operations and help reduce significant societal costs. The key technical challenge to address is problem arises from the fact that CAP-gathered aerial images are spatially sparse and oblique, and well-calibrated object detection datasets are not available for damage-prone situations.

To address the aforementioned challenge, we develop an approach to simultaneously detect and localize damage within images using ideas from weakly-supervised object localization and structure from motion. Firstly, we refine a well-known proposed technique called class activation mapping to detect the extent of damage within an image solely relying on image-level labels. Secondly, we utilize structure from motion to georeference batches of CAP images from an area of interest. The main advantage of our approach is that the outputs of these two techniques can be easily combined to assign real-world coordinates to damage hotspots in the aftermath of a natural disaster. Finally, we evaluate its potential using data from the 2016 Louisiana floods and provide estimates of flood-related damage.

Our approach achieves a precision of 88% when compared against official flooding estimates. Practical deployment of this approach depends on how the current practices and technologies used by CAP are tailored to improve damage detection and localization. To this end, we propose the following technical and policy recommendations: 1) Implement best practices that allow for a high-quality image sequences that can be labeled and georeferenced using modern computer vision techniques; 2) Incor-

porate other sensing modalities such as satellite imagery into CAP imagery analysis for quantitative damage assessment over large spatial regions; and 3) Invest in low altitude imaging technologies and benchmark dataset development.

Thesis Supervisor: Saurabh Amin
Title: Associate Professor of Civil and Environmental Engineering

Acknowledgments

It would have been impossible for myself as an undergraduate senior at MIT to have fully comprehended how emotionally and mentally charged the following two years would be. My time at the Technology and Policy Program started with mass protests that led to the removal of Puerto Rico's governor. It was followed by massive earthquakes in my hometown, the COVID-19 pandemic, protests over racial justice, and an insurrection at the U.S. Capitol, all occurring while I faced tough questions of how I fit into this mess. Frankly I am amazed that I finished this thesis.

All of this to say, this Acknowledgments section should be monumental. This is not just because of the innumerable people that helped me finish this, but also because of how profoundly each of them changed me for the better. It is honestly impossible for me to fit it all into any reasonable length. I will inevitably leave some out. If that is you, know that I am profoundly grateful for having shared my time with you. I hope I can sometime thank you in person.

This line of work would not have succeeded without the aid of my advisor, Saurabh Amin, and of Jeffrey Liu. Your constant feedback, advice and creativity allowed me to take my research past the finish line. Furthermore, you both provided me an opportunity to support younger students through teaching. All of these experiences have helped me grow as a researcher, as a professional and as a leader. I sincerely wish we can continue to collaborate in the future.

This pandemic has been an arduously isolated experience. I would have lost my mind if not for the incessant support of my girlfriend Lexi and of our two cats, Persephone and Freya. You were a constant joy throughout these difficult times and consistently provided advice on how to proceed on matters of life, jobs, research and friendships. I am extremely blessed to have you in my life, and I look forward to our continued shared life in Boston.

I do not think I would have even applied to a master's program if not for the insistence of my parents. Your endless selflessness has inspired me to be the best version of myself. In addition, your unconditional support, along with that of my

brothers Fernando Javier and Gustavo Enrique, has been a source of comfort in challenging times. Thank you for getting me through undergrad, thank you for getting me through graduate school, and thank you for undoubtedly getting me through whatever lies ahead.

Finally, a huge thanks to everyone that helped me get through my time at MIT. This includes my brothers at Theta Delta Chi, my friends from Course 1, TPP, CASE, the incredible folks from the PKG Center, my therapists at MIT Mental Health and Counseling, the amazing staff at Student Support Services, the ARM Coalition, GEL, the brothers of Theta Chi, and many, many others.

To all of you and so many others: from the bottom of my heart, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past two decades, the quantity and impact of extreme weather events has increased at an alarming rate. According to a report by the United Nations Office for Disaster Risk Reduction, between 2000 and 2019 there were 7,348 disaster events that claimed 1.23 million lives and $3 trillion in economic losses. This is a sharp increase from the period between 1980 and 1999, where there were 4,212 events with 1.19 casualties and $1.6 trillion in losses [42]. The cost associated with hurricane damages in 2017 alone, which included Hurricanes Harvey, Irma and María, was estimated at $300 billion [16]. These figures underscore the extent to which responding effectively to natural disasters is a national security priority.

The different stages of disaster response are exemplified in Figure 1-1, which the Federal Emergency Management Agency (FEMA) calls the Recovery Continuum [29]. In the preparedness stage, local and state governments work to educate themselves on the state of the art in responding to disasters, establish protocols to respond to them, and designate roles and responsibilities. In the short-term and intermediate stages, first responders (which now typically includes federal government agencies) work to save people in immediate danger, deliver aid to those who need it, and mitigate damage to the built infrastructure. Finally, in the late-term stage, governments at all stages work to rebuild the physical, social and economic infrastructure in a way that makes it less susceptible to a similar event in the future. The crucial insight from this Recovery Continuum framework is that no one stage is independent of the

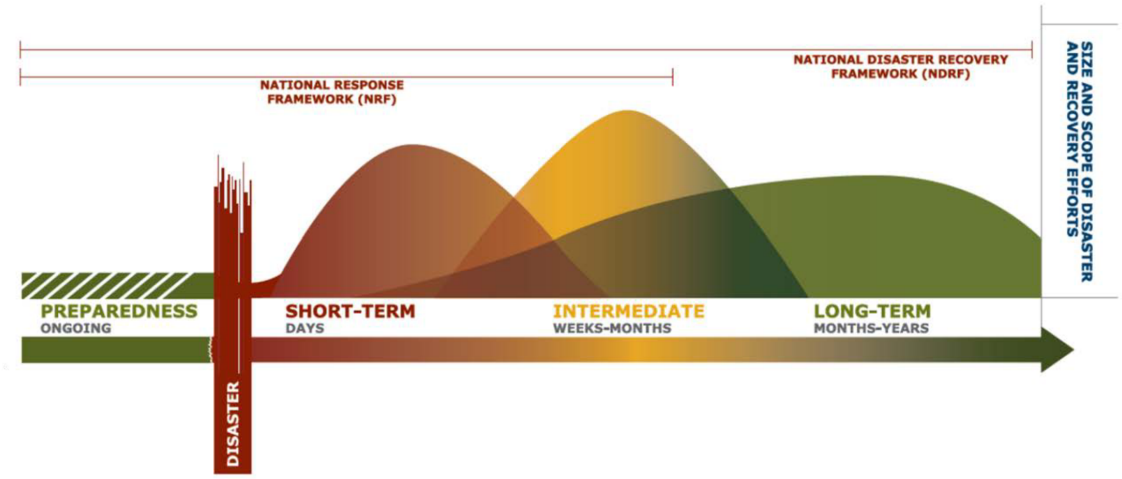


Figure 1-1: FEMA Illustration of the Recovery Continuum [29]

other, but rather work together to mitigate the damage caused by the disasters. Put another way, *it is impossible to have effective disaster response without consideration for all three stages.*

1.1 The short-term recovery stage

This thesis will primarily focus on the short-term stage. While it is certainly just one piece of the puzzle, it presents two main challenges that are not present to the same extent in the other stages. First, the rate at which harm (broadly defined) is done is greater than further out the time horizon. As an example, deaths per day in Puerto Rico between September 20 and 30th, 2017 (immediately after Hurricane María) were an average of 118. Two weeks later, it had fallen to the pre-hurricane average of approximately 80 [36]. Almost by definition, in this stage there is very little time to get it right, and every second saved can dramatically improve the outcome.

Second, there is much more uncertainty about the location and extent of damage. After a disaster, communications networks are typically either overburdened or downright nonoperational. Even after damage is located, the extent of the damage may not be immediately obvious from visual or verbal descriptions. Assessing all sites in person is typically infeasible, as there is a limited amount of equipment and person-

nel to deploy. *Situational awareness* is a precursor to effective disaster response: if you do not know where the damage is, you cannot deploy your resources efficiently. Given the immense data that can be extracted by sensors, automated solutions to processing and analyzing this data are imperative.

Any automated solution towards responding to natural disasters must aid in answering two key questions: *what needs are present*, and *where are these needs spatially located*. The former can include, for example, what types of damage are present in a disaster area, what is the extent or the magnitude of the damage, etc. Meanwhile, the latter can include any estimates of where the resources to attend the needs should be sent. While it is advantageous to have very precise locations for the damage, more often than not the estimates are rather coarse, requiring a team of responders to arrive on the scene and perform further assessment.

1.2 Satellite imagery for disaster response

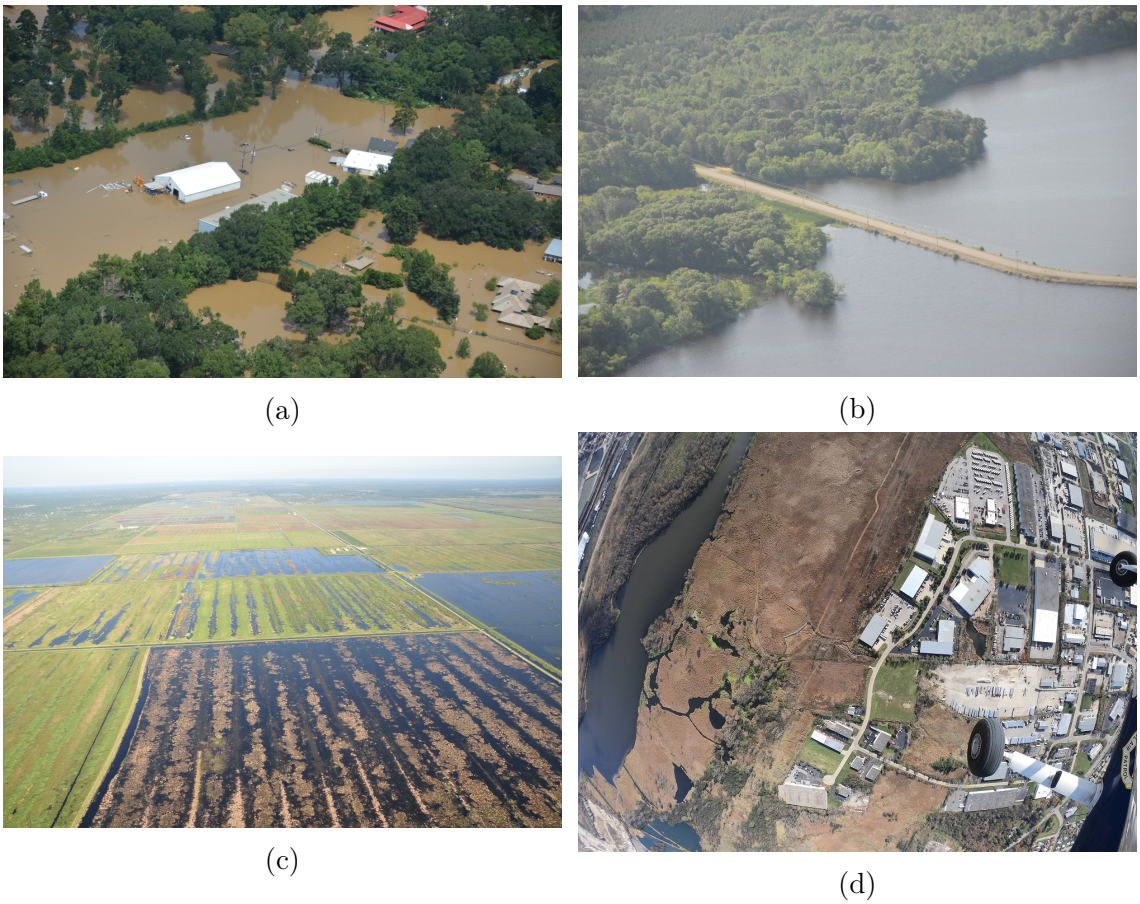
Among the immense data that can be gathered from a variety of sensors, imagery is particularly attractive because it can be easily interpreted by humans. Of the sources of visual data, satellite imagery is the current gold standard for disaster response. Satellite imagery has a number of attractive features. A single satellite image can cover a large area of tens of square miles. The temporal frequency of these images can also be quite high; Planet Labs' Doves, for example, can image most parts of the planet daily [24]. In addition, many satellites have multispectral capabilities, which means they can penetrate certain kinds of occlusion. Finally, satellite images are *georeferenced*, meaning that there is a precise transformation that takes pixel coodinates in the satellite image and converts them to real world coordinates. In terms of the two key questions outlined in the previous section, this means that if you can detect damage on a satellite image, you can exactly determine where that damage is located.

Indeed, there is a fair amount of satellite imagery that is made available specifically for the purpose of responding to natural disasters. Many international space

agencies have agreed to share satellite imagery in response to a disaster situation through the International Charter "Space and Major Disasters" [5]. Private entities have also increasingly allowed access to their imagery for this purpose. Maxar Technologies, which operates the WorldView satellites, makes their imagery freely available through their Open Data Program. Planet Labs, another American remote sensing company, also unofficially supported the International Charter and provides their imagery freely [45].

More than just the images themselves, there has recently been an effort to label images for disaster response. A recent effort is the xView2 challenge, which produced the xBD dataset. This dataset not only labels building footprints, but also provides a label for how much damage the building sustained. As a result of the challenge, machine learning models were created that took the Maxar Open Data Program images and outputted the building footprints and damage labels. These models have already been used in responding to the 2020 Australia wildfires [12].

However, satellite imagery is not without its limitations. To start, satellite imagery cannot have a ground sample distance less than 30cm due to legal restrictions [2]. While this might be appropriate for large scale damage recognition (e.g. buildings), it might not be enough to detect damage to smaller components such as utility poles. Furthermore, while Planet Labs' Doves can image most areas once a day, the frequency is much lower for more powerful satellites. Therefore, images that are of high resolution might not be timely enough to comply with the short-term disaster response necessities. In addition, while many private companies make some of their imagery available, they have no mandate to do so and therefore can choose not to release images that might have some commercial value. Finally and most importantly, while satellite imagery can penetrate some forms of occlusion, it cannot so far penetrate the most common form of occlusion: clouds. In responding to hurricanes in particular, this severely limits the ability of satellite imagery to be helpful in the short-term. Because of these limitations, it is crucial to incorporate data from additional sensing modalities into post-disaster needs assessment.



(a) (b) (c) (d)

Figure 1-2: Sample of CAP images [20].

1.3 Problem statement

One such potential data stream is imagery from the United States Civil Air Patrol. The Civil Air Patrol (CAP) is the civilian auxiliary branch of the United States Air Force. It was created during the Second World War in order to preserve civilian use of aircraft and aid in the war effort [35]. Currently, its functions can be roughly divided in two. First, it provides a variety of cadet and community education programs centered around aviation. Second, it assists with a variety of emergency services [1]. While a significant portion of these are search and rescue missions, the services adapt to whatever the country needs at the time. For example, CAP has delivered medical supplies to areas in need during the COVID-19 crisis [27].

During their post-disaster missions, CAP volunteers take thousands of images of the affected areas. Figure 1-2 shows examples of what these images look like. These

are usually taken from very high resolution DSLR cameras, and are in fact the highest resolution imagery we have available post-disaster for objects "close enough" to the camera. They are also taken just a fairly short time after a disaster event, meaning they could be much more useful for short-term disaster response than satellite imagery can. Furthermore, these images are freely available to all who would want to use them after the fact, making machine learning and other advanced analytics fairly accessible [20].

Given that we have access to high quality, timely and location-tagged imagery, it may be surprising to know that these images are rarely used in practice. This is primarily because interpreting these images is an extremely onerous process. In order to extract information from these images, a human would have to scan through thousands of images for damage and determine where the scene is located. Given the time sensitive nature of a natural disaster, such a process is infeasible. Nevertheless, CAP imagery *ought* to be helpful in increasing situational awareness after a natural disaster. If a human is capable of discerning useful information from an aerial image, it may be possible to automate some or all of this process.

This thesis focuses on the following question: **How can we use modern computer vision techniques to utilize CAP imagery for post-disaster needs assessment, specifically for the purpose of damage estimation and localization?** I broadly define *damage* as an identifiable destruction of utility in infrastructure resulting from a specific event (in our case, a natural disaster). I then define *estimation* as the detection of an instance of damage in an image. Finally, *localization* is the act of assigning world coordinates to the estimated instance of damage. Jointly, I refer to the combination of these two questions as Damage Estimation and Localization (DEL). This is a problem of interest to both Congress and the general public. As a congressionally chartered program, CAP is funded by taxpayers, and therefore there is a government interest that these funds are spent efficiently. Since in the short-term lives are lost at a higher rate, there is also a public interest in ensuring that these services contribute to disaster response capabilities.

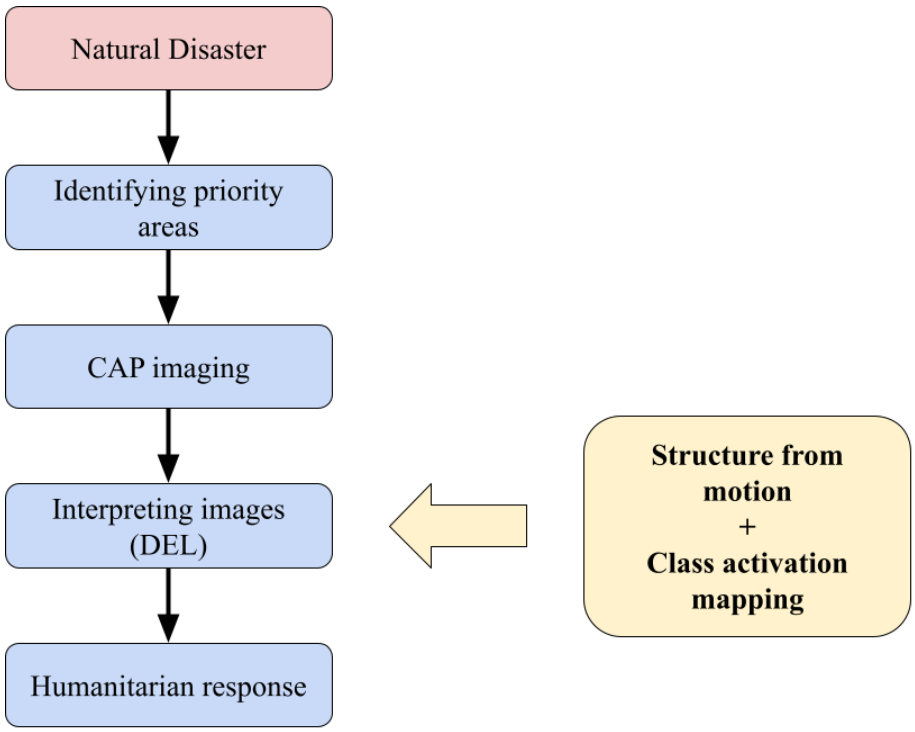


Figure 1-3: Flowchart depicting how CAP imagery would be used in a short-term humanitarian response context. Our approach is highlighted in yellow.

1.4 Contributions

There are three major contributions that stem from this thesis. First, I propose a technical approach towards using CAP imagery to perform DEL. The process of imaging an affected area and acting upon the gathered information is outlined in Figure 1-3. The largest gap in this process is present in the DEL block. Here, I propose an approach that combines two distinct techniques in computer vision to address this DEL gap. The first technique is called *structure from motion*, and it is used to relate coordinates on an image to coordinates in real space. The second technique is called *class activation mapping*, and it is a machine learning approach towards detecting damage within an image. This approach is novel because it helps bridge the gap between the traditional computer vision approaches of the late twentieth century and the modern machine learning based methods.

Second, I examine to what extent CAP imagery is useful as it exists right now in order to inform humanitarian responders on instances of damage. I do this in a

number of ways. To start, I provide a thorough examination of CAP as an entity, its imaging practices, the images that are currently available, and the associated image classification dataset. I then apply the aforementioned DEL approach to a case study consisting of images from the 2016 Louisiana floods in order to compare our flooding estimates with official city data. I perform a rigorous evaluation of the specific metrics under which our approach fares better and worse. Finally, I reason through the performance results using knowledge gained from CAP's organizational characteristics.

Finally, I use the insights from this evaluation to propose interventions that could improve the utility of CAP imagery moving forward. These recommendations span both the technical and policy realms. On the technical side, I highlight specific areas of research that involve incorporating sensors other than CAP into our approach, such as satellite or drone imagery. On the policy side, I urge CAP to document best practices that would maximize the number of images that could be incorporated into our approach; and I urge Congress to invest in CAP imaging infrastructure in a broad sense, ranging from more powerful cameras to improved datasets. The combination of both types of recommendations ensures that CAP is improved in a holistic manner.

The body of this thesis is organized as follows. In Chapter 2, I provide further detail on CAP as an organization, on its imaging program, and on the image classification dataset we use in our analyses. In Chapter 3, I propose technical approach to simultaneously solve both components of the DEL problem. In Chapter 4, I apply the proposed approach to the 2016 Louisiana floods case study, and highlight its current strengths and weaknesses. Finally, in Chapter 5 I outline potential technical and policy interventions that could make the CAP imagery program more effective.

Chapter 2

Organizational Overview and Data Sources

This chapter provides greater detail on Civil Air Patrol (CAP) as an organization, its imaging program, and the Low Altitude Disaster Imagery Dataset. As a technical matter, understanding the context in which aerial disaster imagery is taken helps inform us which tools are most appropriate in order to incorporate these images into disaster response. Furthermore, this helps us reason through which policy interventions could be more or less effective in order to improve the imaging program. As we see in Chapter 4, the effectiveness of using CAP imagery is constrained by the processes that shape CAP itself as well as our image classification dataset. This chapter will identify what the key features of these processes are.

2.1 Overview of the Civil Air Patrol

As mentioned previously, CAP is the civilian auxiliary branch of the United States Air Force. It is comprised almost entirely of volunteers, who pay membership fees and other expenses. Only more senior administrative members of CAP receive any sort of pay for their work [32]. While CAP is an auxiliary of the United States Air Force, military officers do not have authority over CAP members. The United States Air Force can request assistance from CAP for non-combat missions, but CAP is

typically reimbursed for any expenses associated with these missions. Furthermore, CAP is only deemed an instrumentality of the Untied States when carrying out a mission assigned by the Secretary of the Air Force [1].

CAP is divided into eight regions corresponding to various geographies in the United States (*e.g.* Northeast Region and Pacific Region), which are then further divided into wings for each state and territory. Each wing is comprised of various units corresponding to the different localities within the state or territory [32]. While there are processes and procedures that are intended to be followed throughout all of CAP, the level of enforcement varies. In terms of disaster response, the most local units are slated to respond first, followed by the state wing and if needed the region [32]. Different regions can collaborate on the same emergency response. This can be the case either because an emergency spans multiple CAP regions (such as the COVID-19 crisis), or because the FEMA response regions are misaligned with the CAP regions.

CAP does not respond to a disaster situation by its own accord. Rather, it is tasked by some other emergency response entity (such as FEMA, the Department of Homeland Security or a local fire department) with specific goals, such as search and rescue, aerial transportation or damage assessment. Since CAP is operated by volunteers, hiring CAP for these missions is much less expensive than hiring a private entity. CAP solely charges for the expected cost of carrying out the mission, such as fuel and maintenance, which is typically $160-200 per hour [32].

Given that this thesis revolves around CAP imagery, it is important to highlight some key features of the imaging process. Every CAP aircraft is equipped with a NIKON Pro Quality camera. Some models used are the NIKON D90, D5100 and D7100 [32]. These are handheld cameras that are intended to be used by a photographer on the passenger seat to image targets requested by the client. As a result, these images are almost always oblique (*i.e.*, at an angle relative to the ground), as shown in Figure 1-2. The vast majority of these images are geotagged (meaning that they have a rough estimate of the camera's GPS position at the time the image was taken) and timestamped. From looking at the timestamps, we see that the vast majority

of images were taken in the past five years. This likely corresponds to a number of factors, such as the increased affordability of high quality cameras and the launch of the online CAP imagery portal. It is fairly likely that CAP volunteers around the country took images in the 2000's that were never uploaded.

CAP offers a very comprehensive Airborne Photographer Task Guide that outlines best practices for imaging targets. It includes general guidelines on how to operate the camera, how to take effective photos, how to plan an imaging sortie, and how to upload the images to the portal [31]. While these guidelines are available, it is not clear how much photographers adhere to these best practices.

An important limitation of this mode of imaging is that, while the images are geotagged they are not *georeferenced*. As mentioned previously, georeferencing refers to finding a linear transformation that relates image coordinates and scene coordinates. In short, it shows where the contents of the image are located in real coordinates, as opposed to the geotag which shows where the camera is located. Since these images are taken at an angle, the geotag is usually not a good approximation to the contents of the image. Clearly the location of the scene is much more important for effective disaster response than the location of the camera. While it is possible to georeference images taken from handheld cameras (as discussed in Chapter 3), this is only possible under very specific imaging conditions. These conditions are not included in the given Task Guide.

CAP has started taking steps to address this limitation through the introduction of two technologies. One of these is the Garmin VIRB camera, which provides top-down images when attached to the wing of the aircraft. The other is the WaldoAir XCAM sensor, a relatively low cost multispectral imaging system. Both of these allow CAP to perform automatic georeferencing of the surveyed area. They are currently being tested by CAP as a suitable addition to their aircrafts [15]. However, rolling out these powerful sensors more broadly is not likely to happen in the near-term. This is because these sensors are much more expensive than the NIKON cameras, and because of existing best practices associated with the current process. Thus, locating a scene within an image remains a challenge.

2.2 Low Altitude Disaster Imagery Dataset

Another challenge with working with CAP imagery is interpreting them. Indeed, locating the contents of an image is not helpful if we cannot understand what is in the image. Modern approaches in a variety of contexts rely on machine learning, whereby a model is trained to detect objects in images by detecting patterns in training datasets with ground truth annotations. Until recently, such datasets did not exist for this context.

In 2019, Liu et al published the Low Altitude Disaster Imagery (LADI) dataset [20]. LADI is a publicly available dataset consisting of images taken by CAP in the aftermath of natural disasters, and annotated by crowdsourced workers with hierarchical image-level labels representing five broad categories: Damage, Environment, Infrastructure, Vehicles, and Water. Within each category, there are a number of more specific annotation labels. Table 2.1 shows the available categories and labels for the LADI dataset. The current LADI dataset does not provide any bounding box or segmentation information.

Damage	**Environment**	**Infrastructure**	**Vehicles**	**Water**
damage (misc)	dirt	bridge	aircraft	flooding
flooding / water damage	grass	building	boat	lake / pond
landslide	lava	dam / levee	car	ocean
road washout	rocks	pipes	truck	puddle
rubble / debris	sand	utility or power lines / electric towers		river / stream
smoke / fire	shrubs	railway		
	snow / ice	road		
	trees	water tower		
		wireless / radio communication towers		

Table 2.1: Available categories (in bold) and labels for the LADI dataset.

The LADI dataset was generated using Amazon Mechanical Turk, an online platform where anonymous workers can fulfill preset tasks. In this case, a variable number

of workers were paid to classify an image as one or more of the annotations presented in Table 2.1. While these workers were not experts in disaster response necessarily, some filtering or workers was done to ensure a level of quality. Workers were initially evaluated on their ability to recognize annotations from a small subset of CAP images.

To my knowledge, this is the only freely available dataset that consists of post-disaster images taken from oblique angles. As alluded to in Chapter 1, the vast majority of work in this space has revolved around orthorectified imagery from satellites or aerial platforms [12, 6, 34].The only similar dataset I encountered was the Volan2018 dataset, which provides bounding box information for disaster imagery for various related classes [33]; however, at the time of writing, it is not publicly available. Because of this, I rely on the LADI dataset for the analyses in this thesis. For the reader's benefit, in Appendix A I included a list of similar datasets with short descriptions.

2.3 Practical bottlenecks

Given our goal of using CAP images to perform DEL, it is important to outline key challenges in operationalizing these images. One set of challenges revolves around the way that the images were taken. Specifically, the fact that these images are oblique means that we cannot georeference them relative to an orthorectified satellite or aerial image. In the case of satellite or top-down drone imagery, others have been successful in applying visual feature-based methods such as SIFT to register the image in question to another whose geotransform was known [22, 30, 48, 11]. However, methods such as SIFT have been shown to perform poorly under extreme changes in perspective and sensor specifications [48, 37], as was the case when I attempted to use SIFT to georeference post-disaster aerial images and satellite images. Previous research also attempted to geolocate images using Siamese neural networks in one of two ways: either training the network to match certain features (*e.g.* buildings) of a query image and a ground truth image [40] or by matching the entirety of a

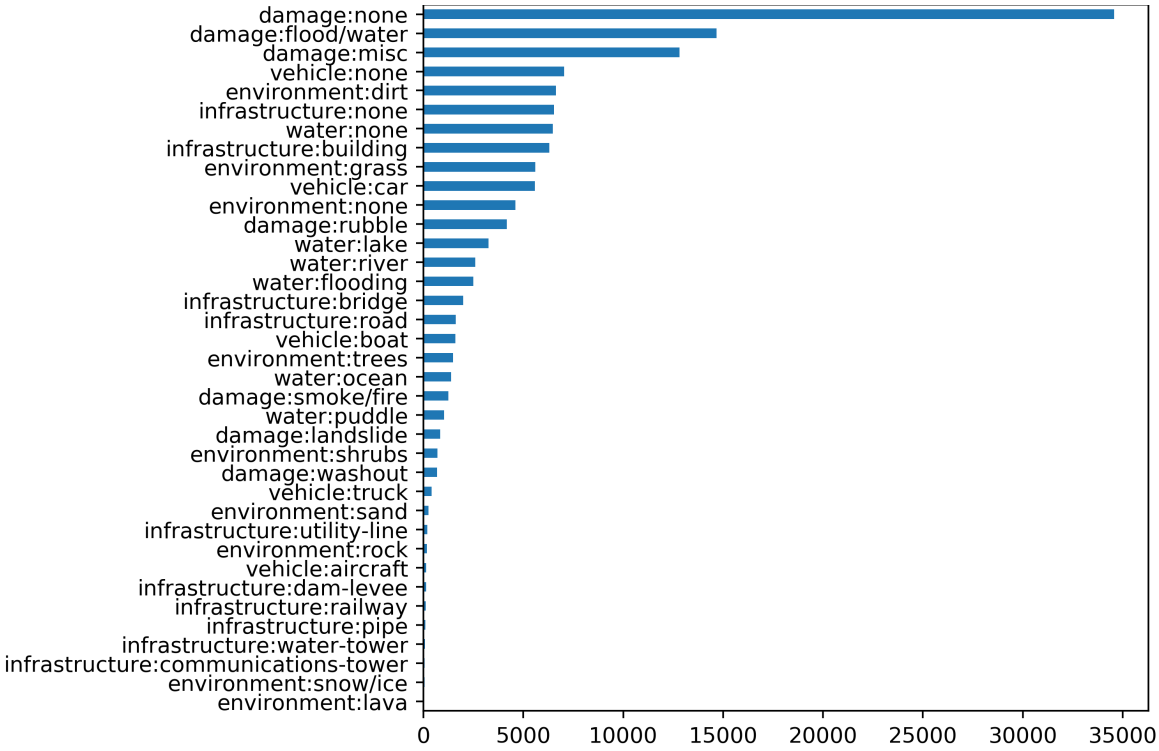


Figure 2-1: Bar plot showing the amount of images for which at least one worker classified it as a particular label.

query image and a ground truth image [17, 37, 21]. While these approaches would be suitable for estimating the GPS tag of aerial images, estimating the full geotransform would require additional orientation information.

The other major challenge revolves around the labels that are available for this type of imagery. Given that the LADI dataset is the only freely available dataset in this context, we do not have access to popular image segmentation or object detection techniques. Furthermore these labels are not equally represented in the dataset. Figure 2-1 shows the amount of times each label was identified within an image by at least one worker. I reason that the disparity in the labelling is due to two factors. First, some labels are more easily identifiable than others by non-experts, especially in certain contexts. For example, while landslide in an urban center might be easily spotted, rubble in a more mountainous region might not be seen by non-experts as damage. Second, some labels are simply imaged less. For example, in the United States lava is only present in Hawaii, where CAP availability is more limited. Other

labels, such as smoke and fire, represent flying hazards and are thus unlikely to be captured by manned aircraft.

These two factors point to sources of bias within the available dataset. Indeed, what types of damage are imaged and what damage is recognized in an image are points in the data collection process where the biases of the client, of the CAP volunteers and of the labellers come into effect. Understanding the dynamics that produced this dataset can hint at which scenarios we can expect to do a reasonable job at detecting damage. Having taken these factors into consideration, I decided to limit our analyses to detecting and locating flooding from LADI images. Of the damage labels, Figure 2-1 suggests that flooding is likely to be the one we should expect the best performance on.

Now that we have a full understanding of CAP's disaster imaging program and of the LADI dataset, our goal is to use CAP imagery to perform DEL. In order to do so, we need to work around the two major sets of challenges outlined above. In the following chapter, I will incorporate these insights in order to propose a technical approach towards solving this problem.

Chapter 3

Damage Estimation and Localization

3.1 Description of the technical approach

This chapter is devoted to outlining our technical approach towards using CAP imagery to perform damage estimation and localization, or DEL. As mentioned in Chapter 2, given the state of the current LADI dataset I have decided to focus solely on detecting and localizing flooding. However, it is important to emphasize that the methods discussed here are much more broad, and can be generalized to any DEL of any other type of damage.

Figure 3-1 provides a visual depiction of our approach. It consists of two stages: a pre-disaster and a post-disaster stage. In the pre-disaster stage, a neural network is trained to recognize image-level damage labels within an aerial disaster imagery dataset. The post-disaster stage is comprised of two parallel pipelines, whose outputs are combined at the end. The first pipeline takes a collection of images from an area of interest and reconstructs the scene using structure from motion. The reconstructed point cloud then relates image coordinates to world coordinates via a projective transformation. The second pipeline takes individual images from the area of interest and produces polygons that cover the extent of the damage that is detected using class activation mapping. Finally, the projective transformation is applied to the damage polygons to produce the DEL output. While our approach uses image-level binary damage indicators and class activation maps due to the limited

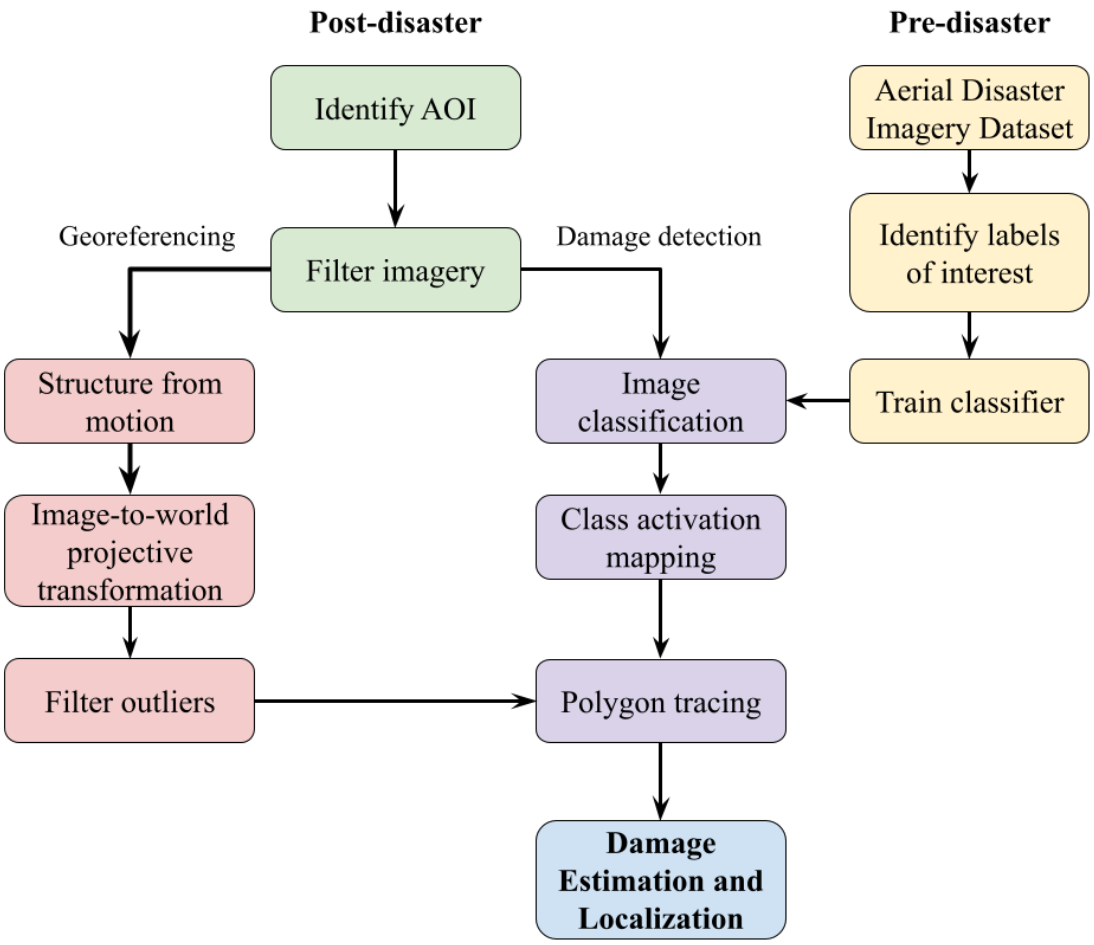


Figure 3-1: Flowchart depicting our approach. The different colors represent different components that can be done independently.

availability of training data, the same estimated projective transformation could be applied to bounding boxes or segmentation masks generated from object detection or semantic segmentation algorithms, respectively.

This chapter is organized as follows. Section 3.2 describes how I use image classification and class activation mapping to estimate the extent of flooding in an image. Section 3.3 describes how I use structure from motion to reconstruct portions of the area of interest and estimate a projective transformation relating image and world coordinates, which is then applied to the class activation maps in order to finalize the DEL.

3.2 Damage estimation using class activation mapping

This section describes our damage estimation pipeline. Given an image, our goal is to find the extent of flooding within an image. The main technique used here is called *class activation mapping*. Put simply, it takes a neural network that has been trained on classifying images and produces a map of what areas of an image are most related to a specific class. Because of this, I initially pose this as a classification problem of detecting damage within an image. Then, I use class activation mapping to find the boundaries of damage within the image.

3.2.1 Review of image classification

Image classification is a specific task in computer vision where a set of images is given one or more classes. The goal is then to determine the image's true class. Modern approaches use a *supervised learning* paradigm. This paradigm is illustrated in Figure 3-2. Assuming you have "enough" images with known labels, these images are passed through some machine learning model that learns underlying patterns related to a specific class. Once that model has been sufficiently trained, it is then able to detect those same patterns on new images in order to classify them.

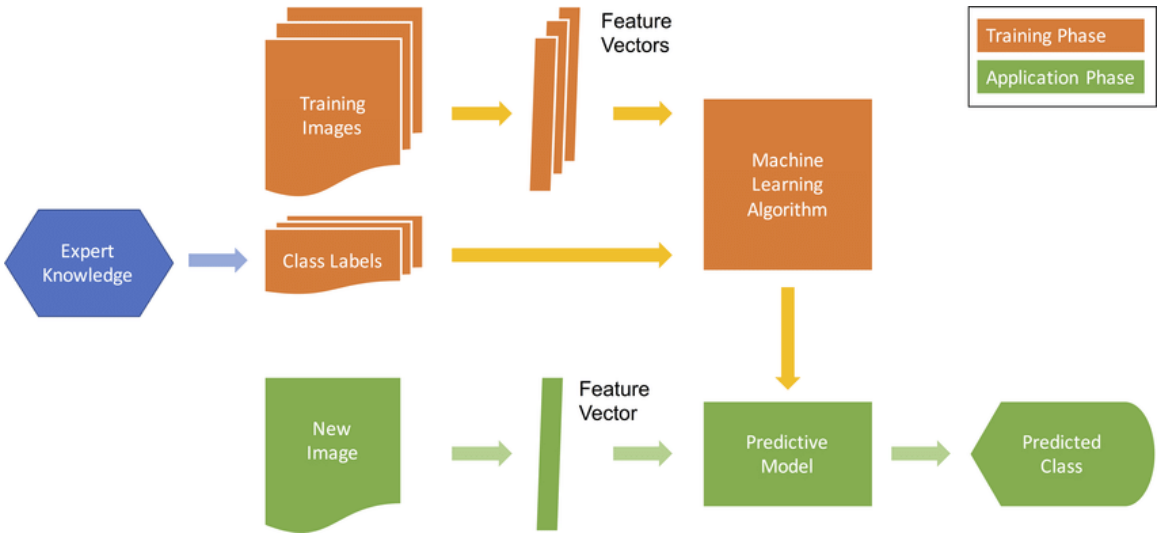


Figure 3-2: Steps in a supervised learning pipeline [43].

The specific machine learning model that is deployed depends on the context. An important class of machine learning models are deep neural networks. These are models comprised of various layers, which are themselves made up of units. The connections between successive units and layers encodes over the period of training the relative importance of particular features towards a specific class. Specifically for image classification, *convolutional neural networks* (CNN) have been especially popular. First introduced by LeCun et al in 1989 [19], CNNs include convolutional layers that are appropriate for detecting visual features such as corners or edges. CNNs have proven to work exceedingly well in a variety of image classification tasks [18, 14, 38]. Therefore, they are an appropriate first step in our DEL problem.

3.2.2 CNN implementation details

Our first step is to detect whether an image contains flooding anywhere in the image. I pose this problem as a classification problem, where for an image X_i there is a label $Y_i^{true} \in \{0, 1\}$ that corresponds to whether an image contains flooding or not. Our goal is to predict Y_i^{true}. In the construction of the LADI dataset, images were shown to a variable number of workers (generally between 3-5); each worker was asked to identify which, if any, labels for a given category (*e.g.* "Damage") applied to that

image [20]. Responses from all workers were recorded. As a result, images could have potentially differing annotations between workers due to factors such as subjectivity (worker did not think the label applied) or error (incorrect input). To account for these conflicts, I designate three different labelling schemes for classifier training and evaluation:

A) $B_{i,j} > 1$,

B) $B_{i,j} > 2$,

C) $B_{i,j} > 1$ and $B_{i,j}/w_i > \underset{\forall i}{\text{median}}\{B_{i,j}/w_i\}$,

where $B_{i,j}$ is the number of workers that labelled image i as class j and w_i is the number of workers that labelled image i at all. Training and testing using different combinations of these labelling schemes ensures that we achieve a balance between filtering out noise and preserving a sufficiently representative to train on.

To perform the image labeling task, I use ResNet-50 as a backbone architecture due to its impressive performance in a variety of image classification tasks [14]. ResNet-50 is a CNN that is 50 layers deep. The main innovation of the ResNet architecture compared to other feed-forward CNN architectures such as VGG and AlexNet [18, 38] is the inclusion of shortcut connections that skip subsequent layers and perform identity mapping. The ResNet architecture has been shown to solve the degradation problem [13], whereby increasing the depth of a neural network results in saturation and eventually degradation of its accuracy.

I split the dataset 80%/10%/10% for the training, validation and testing sets, respectively. Images were scaled such that the shorter dimension was 224 pixels long, and then cropped into a 224×224 tile. Random rotations and horizontal flips were applied during training. I initialized the ResNet with weights pretrained on the ImageNet training set [7], and changed the output layer dimension from 1000 to 1. I then trained the CNN with a batch size of 8, a learning rate of 0.001, a momentum of 0.9, using stochastic gradient descent as the optimization algorithm with the Binary Cross-Entropy (BCE) loss:

$$L = \sum_{i=1}^{m} Y_i^{true} \log \sigma(Y_i^{pred}) + (1 - Y_i^{true}) \log \sigma(1 - Y_i^{pred}), \tag{3.1}$$

where $Y_i^{pred} \in \mathbf{R}^1$ is the output of the neural network, m is the number of images in the batch and $\sigma()$ represents the Sigmoid function.

3.2.3 Class activation mapping and polygon tracing

After training, I follow the class activation mapping (CAM) approach from Zhou et al [47] to localize the extent of the detected class within the image. CAM is a technique for weakly-supervised object detection (*i.e.*, where bounding boxes are not explicitly trained on) which takes advantage of the average pooling layer at the end of the ResNet architecture to detect areas within an image that are important for classifying a particular class. Using the terminology in [47], the output on the final fully connected layer is given by:

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y), \tag{3.2}$$

where w_k^c is the weight corresponding to class c for the kth unit within conv5 (the final block within ResNet-50) and $f_k(x, y)$ is the activation of the same kth unit at location (x, y) (such that $\sum_{x,y} f_k(x, y)$ is the output of the global pooling layer). Let $M_c(x, y) = \sum_k w_k^c f_k(x, y)$, so that:

$$S_c = \sum_{x,y} M_c(x, y) \tag{3.3}$$

Here, $M_c(x, y)$ corresponds to a measure of importance of a spatial coordinate (x, y) for the class $c =$ *flooding/water damage*, which corresponds to the class activation map. In order to determine the boundaries of the flooding instances, I perform a simple thresholding on M_c:

(a) Original image (b) CAM mask (c) Polygon tracing

Figure 3-3: Stages of our polygon tracing approach.

$$M_c^{\text{mask}}(x, y) = \begin{cases} 1 & \text{if } M_c(x, y) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

The last step in the estimation pipeline is to convert the masked image into a set of polygons, so that I can easily transform the boundaries of flooding across coordinate systems. I do so by first extracting countours using [39], which are then categorized into parent (the outer border of a component) and child (the border of any holes within a component) components. Polygons are then saved from these parent/child components, so that the projective transformation relating image and world coordinates can be applied to them. Figure 3-3 shows the different stages of this procedure.

3.3 Damage localization using structure from motion

Our goal now is to transform these polygons, which are in image coordinates, into world coordinates, which is our localization pipeline. The main technique used here is called *structure from motion*, which essentially finds common point within images and uses known intrinsic camera parameters to estimate depth in an image. The process of using structure from motion to estimate the projective transformation relating image

and world coordinates and applying the transformation to the flooding polygons is described in this section.

3.3.1 Review of structure from motion

Structure from motion is a technique that, using images from a camera moving through an environment, can produce a point cloud of the environment [3]. By taking advantage of the GPS tags from the image metadata or from outside sensors, structure from motion has been used to create inexpensive, georeferenced elevation models from drone and aircraft imagery [9]. I take this same approach as an intermediate step to obtaining the projective transformation that relates image coordinates to world coordinates.

Suppose you have two images from two different pinhole cameras with two centers of projection, C_1 and C_2, and in these images you have a series of corresponding points (which are points on the image that are of the same object). I denote those points $m_{i,j}$, where:

$$m_{i,j} = \begin{bmatrix} x_{i,j} \\ y_{i,j} \\ 1 \end{bmatrix} \tag{3.5}$$

Here, $x_{i,j}$ and $y_{i,j}$ are the pixel coordinates of point i in image j. These coordinates are measured from the top left corner of the image. The relationship between these points is shown in Figure 3-4. Note that a single image is not sufficient to uniquely determine the position of M, as it could be anywhere on the ray connecting the center of projection (in the diagram, C_1 and C_2) to the point on the image. Conversely, if we have two images with a corresponding pair of points *and* the camera pose (rotation and translation) then we can uniquely determine the location of M.

Now suppose that the two cameras are calibrated. This means that we know the principal point (the intersection of the perpendicular ray that connects the center of projection and the image plane) and the focal length (the distance in pixel length between the center of projection and the image plane). We can express this in matrix

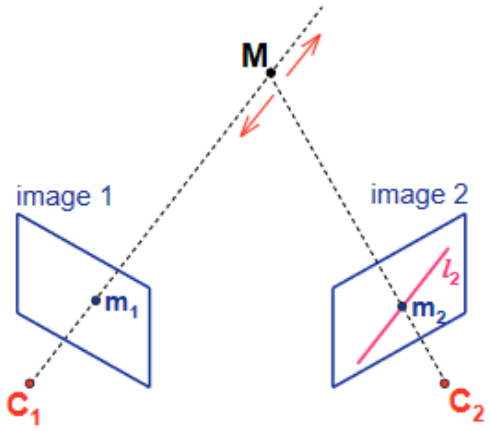


Figure 3-4: Diagram showing two corresponding points, m_1 and m_2, of a real world point M [25].

form as:

$$K = \begin{bmatrix} \alpha, 0, p_x \\ 0, \alpha, p_y \\ 0, 0, 1 \end{bmatrix} \tag{3.6}$$

Here, α is the focal length, and (p_x, p_y) are the coordinates of the principal point. We can now relate the two sets of corresponding points by the *essential matrix* E:

$$(K^{-1}m_{i,2})^T E K^{-1} m_{i,1} = 0 \tag{3.7}$$

This essential matrix has very desirable properties. First, it can be estimated using just four corresponding pairs of points. This estimate can be made more robust by adding more pairs. Second, it can be decomposed into a rotation matrix R and a translation vector t up to a scale factor. Given that we are provided GPS tags for all the images, we can recover the scale factor and arrive at a unique solution [25]. From an initial reconstruction, additional images can be added either by repeating the same process or by a method called *Perspective-n-Point*, where the camera pose is automatically estimated by linking corresponding points in the new image with points that have already been reconstructed in 3D.

3.3.2 Reconstruction using OpenSfM

I base my implementation off the OpenSfM library, an open source library for structure from motion [23]. I first choose a specific area of interest and filter out all images within the LADI dataset whose GPS tags were not within 5 km for the area of interest boundary. Following the filtering step, HAHOG features (which are a combination of the Hessian affine feature point detector and the histogram of oriented gradients descriptor) are extracted from all images and matched across images using the FLANN algorithm [26]. Then, a reconstruction is initialized from two views by estimating the essential matrix and decomposing the matrix into translation and rotation components [3]. Afterwards, additional images are added to the reconstruction using Perspective-n-Point [8]. After each image is added, the camera poses and reconstructed 3D points are jointly optimized (a process called bundle adjustment [41]). OpenSfM then uses the GPS tags to properly align the reconstruction in world coordinates. If images still remain that have not been added to an existing reconstruction, a new reconstruction is initialized and the process repeats until no more images can be reconstructed. Finally, OpenSfM outputs a collection of camera poses and reconstructed feature points, in east, north, up (ENU) coordinates with the average of the GPS coordinates as the reference point. ENU coordinates are measured in meters from a reference point (in our case, the average of the GPS tags), and are aligned so that the x-, y- and z-axis are aligned with the east, north and up directions, respectively.

3.3.3 Estimating the up-vector

Because fixed wing aircraft have a relatively large turning radius compared to rotary-wing aircraft, sequential images collected from fixed-wing platforms tend to be approximately collinear. This means that some reconstructions potentially have an additional degree of freedom from rotating about the line that goes through the GPS coordinates. Therefore, it is necessary to estimate the direction of the up-vector (*i.e.*, the vector opposite to the direction of gravity) and enforce it in the reconstruction.

Previous implementations of structure from motion in urban environments have suggested estimating vanishing points to estimate the up-vector [44]. This can be difficult if there are few straight features, such as roads, or high amounts of vegetation, which is common in rural areas.

To address the issue of estimating the up-vector, I propose an approach which assumes the ground is approximately flat. I first fit a plane through the reconstructed features using RANSAC [8]. There is a pair of possible antiparallel unit normal vectors to this plane, one of which is the up-vector. Because of the aerial nature of the data, the location of the images must be above the ground plane. Therefore, I choose the vector that has a positive projection onto the image location in ENU coordinates and denote it v_{up}. Finally, I rotate the reconstruction so that v_{up} indeed points upwards. Specifically, I rotate it by R_z such that $R_z v_{up} = \hat{z}$ when it is initialized, and the up-vector is enforced during bundle adjustment.

3.3.4 Image-to-world projective transformation

The final step in our georeferencing pipeline is estimating the transformation from image coordinates to world coordinates and applying this transformation to the detected damage polygons. As discussed previously, the images are of mostly flat surfaces, meaning both sets of coordinates can be related by a projective transformation that can be estimated with at least four correspondences [3], and outliers can be filtered through RANSAC [8]. Of all of the images that were reconstructed using OpenSfM, I retained those where at least 20% of matches between image coordinates and world coordinates were inliers.

Of the retained images, I found that some images produced extremely large image footprints (*i.e.*, the projection of the image edges onto the ground). Upon inspection, I saw that these were images that were so oblique that the horizon was visible. Because these images require more complex transformations, I decided to disregard these images for our implementation. I considered two criteria for eliminating such images. First, I only eliminated images whose total area were greater than some value γ_1. Second, I did not consider images where the ratio of the longest side to the shortest

side of the minimum area rectangle that covered the entire footprint were greater than γ_2. The projective transformation is applied to all polygons generated by the procedure in Section 3.2 to obtain our flooding estimate. I chose the values of γ_1 and γ_2 empirically to maximize the prediction precision.

This projective transformation is the link between the structure from motion and the damage detection branches in Figure 3-1. Thus, applying this projective transformation allows us to automatically interpret CAP images and perform DEL. Of course, the actual performance of this method remains to be evaluated. This evaluation is the focus of the following chapter.

Chapter 4

Evaluation on a Case Study: the 2016 Louisiana Floods

In this section, I provide an evaluation of my approach at predicting real-world instances of flooding. An ideal evaluation would involve comparing the output of my structure from motion and class activation mapping pipelines against ground truth values of the geotransform and the extent of flooding of an image, respectively. In order to quantify such an evaluation, I would present the intersection over union metric for both the transform of the CAP image and the flooding polygons within the image. However, neither of the required data are available. Given this, both of the ideal evaluation methods are infeasible. The most I can present are qualitative evaluations of whether my approach seems reasonable.

Instead of the ideal quantitative evaluation, in this thesis I present an evaluation against data from the East Baton Rouge parish of the 2016 Louisiana Floods. Figure 4-1 shows the administrative boundary of the East Baton Rouge parish in Louisiana, the parish's estimated flood inundation area [28], and the coordinates of all CAP image with GPS locations within 5 km of the administrative boundary. In total, the flooding event covered 536 km^2 (44% of the total area of the parish). My analysis includes 1615 CAP images that were taken in August 2016 immediately after the flooding event. Comparing the output of my approach against this data can serve as proxy for having the geotransforms and the segmented images.

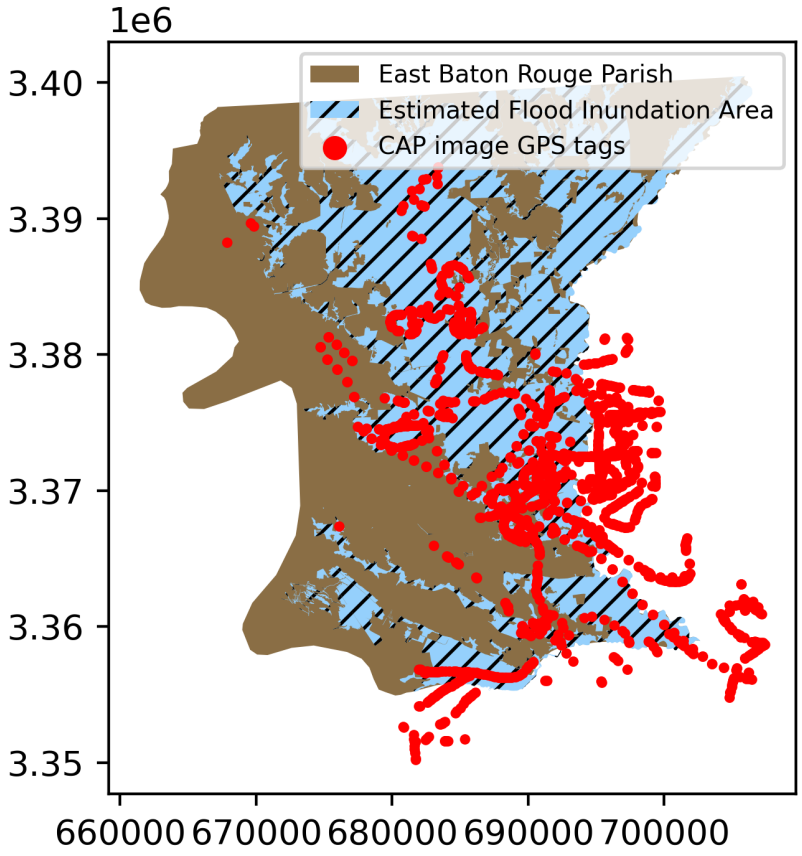


Figure 4-1: Map of East Baton Rouge parish from the 2016 Louisiana floods, along with the GPS tags of all images taken within 5 km of the parish boundary. Map uses UTM 15N coordinates.

This chapter is organized as follows. In Section 4.1, I evaluate the performance of my image classifier for the *flooding/water damage* label on the entire LADI dataset, using the different labelling schemes described in Section 3.2.2. Next, in Section 4.2, I present the results of the class activation mapping procedure for localizing *flooding/water damage* within an image. In Section 4.3, I apply my approach to the 2016 Louisiana Floods data to classify and localize flooding damage captured from CAP photographers, and compare it against the parish's post-event estimates of the flooding extent to evaluate the precision of my approach. Finally, in Section 4.4 I discuss some key considerations regarding extending my approach to other labels.

Train label	Test label	Accuracy (%)	Precision (%)	Recall (%)
A	A	77	65	79
	B	71	38	91
	C	70	42	83
B	A	77	75	53
	B	83	54	68
	C	80	55	64
C	A	79	73	65
	B	80	48	80
	C	83	53	69

Table 4.1: Accuracy, precision and recall values for the three ResNet models. *Train label* refers to the set of labels used in training, while *Test label* refers to the set of labels against which each model was evaluated.

4.1 Classification results

Table 4.1 shows the testing accuracy, precision and recall values for the three ResNet50 classifiers that were trained (one for each ground truth training labelling scheme defined in 3.2.2). In order to properly compare the three models, each of the three classifiers was also evaluated against the remaining two labelling schemes. Regardless of the labelling, the actual images that comprised the testing set (as well as the training and validation sets) were the same for all three schemes. For the purposes of this thesis, I will refer to each of the models according to their training labelling scheme.

Unsurprisingly, each of three models had the highest accuracy when compared against the labelling scheme they were trained on. With the other two metrics, though, there are noticeable trends. In terms of precision, model B had the highest precision when evaluated against any labelling scheme, followed by C and finally A. With recall, the opposite holds: A has the highest recall across the board, followed by C and then B. These trends are not difficult to justify, since B necessarily has a higher standard for classification as flooding than A. For flooding, C is a compromise between the most lenient labelling scheme and the strictest one. In the particular context of disaster response, we consider false positives to be of lower regret than false negatives. As such, I proceed using model A for the remainder of the section.

4.2 Class activation mapping results

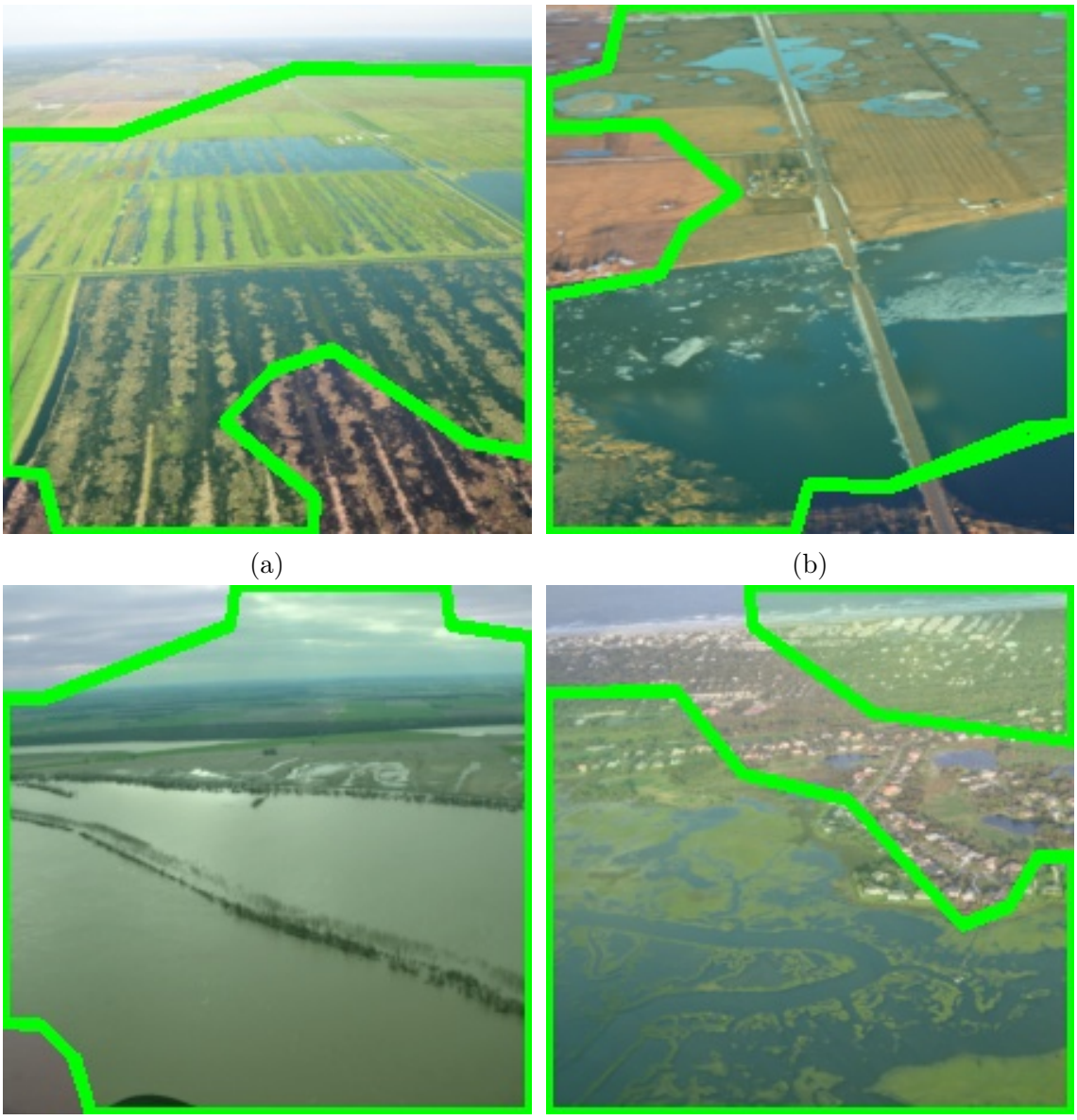
I now present an evaluation of the class activation mapping procedure on the flooding images. While my observations in this section will be qualitative, in Section 4.3 I will quantitatively evaluate the performance of my approach derived from the class activation map as a proxy for evaluating the efficacy of the class activation mapping procedure.

Figure 4-2 shows a sample of LADI images that were classified as having *flooding/water damage*, along with the estimated extent of flooding. We can see that for the most part, this method does an adequate job of tracing the extent of water in the image. Flooding is slightly more complicated. While Figures 4-2a, 4-2b and 4-2c very clearly show flooding events, the top portion of Figure 4-2d seems to simply be picking up the shoreline. As an important note, I noticed that many images that show large bodies of water also tend to include the horizon in the flooding polygon (*e.g.* Figure 4-2c). This might be because flooding typically covers a large portion of area, and therefore images that include the horizon might be more likely to also include flooding. This underscores the importance of filtering images with large footprints after georeferencing. Figure 4-3 shows images that were identified as flooding from the Louisiana 2016 floods. Even in this case where many images have large portions of flooding, my approach is still able to trace the extent of the water.

4.3 DEL results

Of the 1615 CAP images that were considered, 809 were successfully reconstructed by OpenSfM. At the same time, of the 1615 images 996 were identified as having flooding. Finally, 559 images completed the georeferencing pipeline *and* were identified as flooding. Additional images were the filtered based on the criteria described in Section 3.3.4.

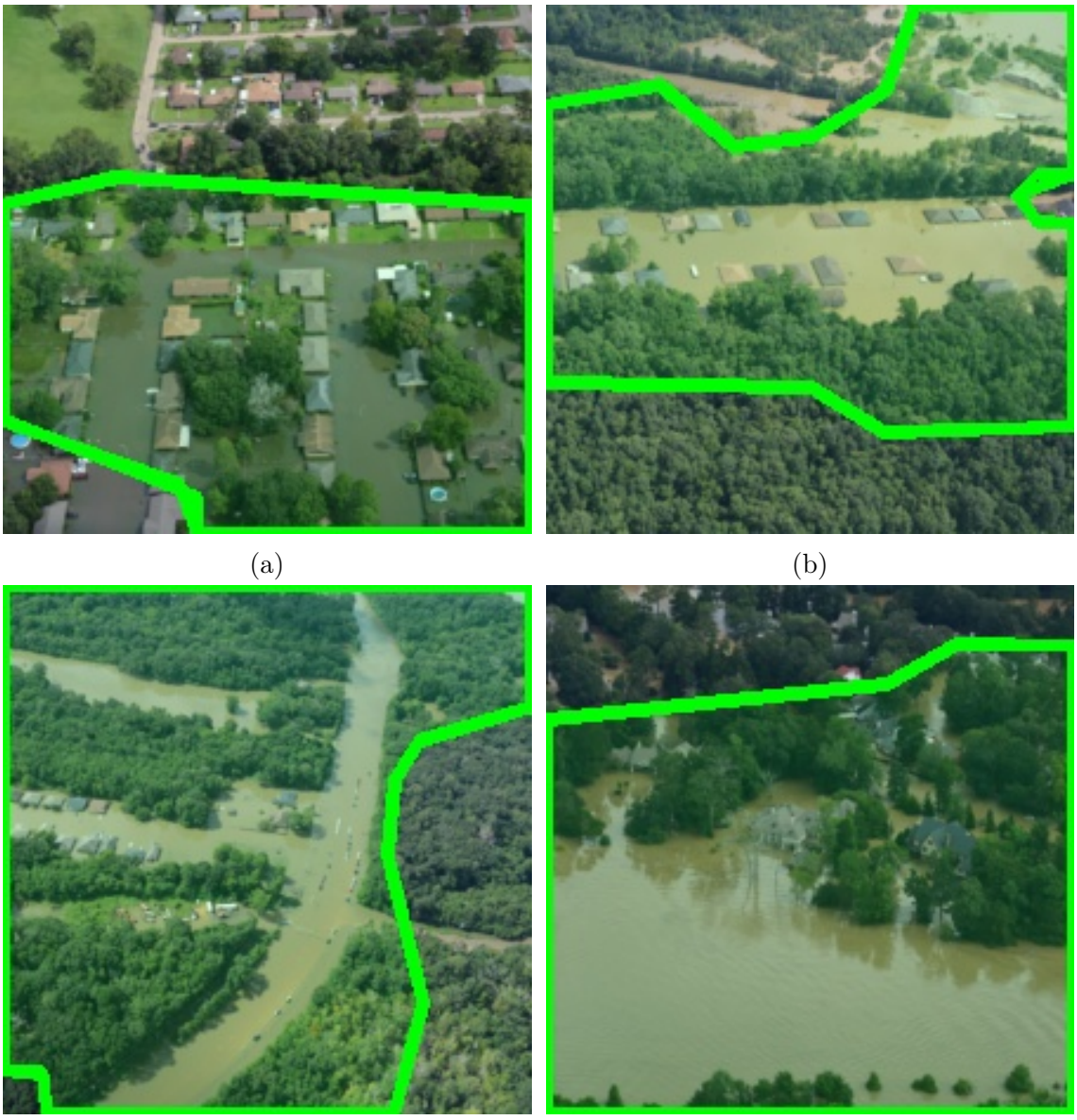
I used these images to estimate flooding using three approaches. First, I use the GPS tag of these images as a baseline, where I calculate the precision as the proportion



(a) (b)

(c) (d)

Figure 4-2: Sample of CAP images identified as flooding by ResNet model A and their associated damage polygons.



(a) (b)

(c) (d)

Figure 4-3: Sample of CAP images from the 2016 Louisiana floods and their associated damage polygons.

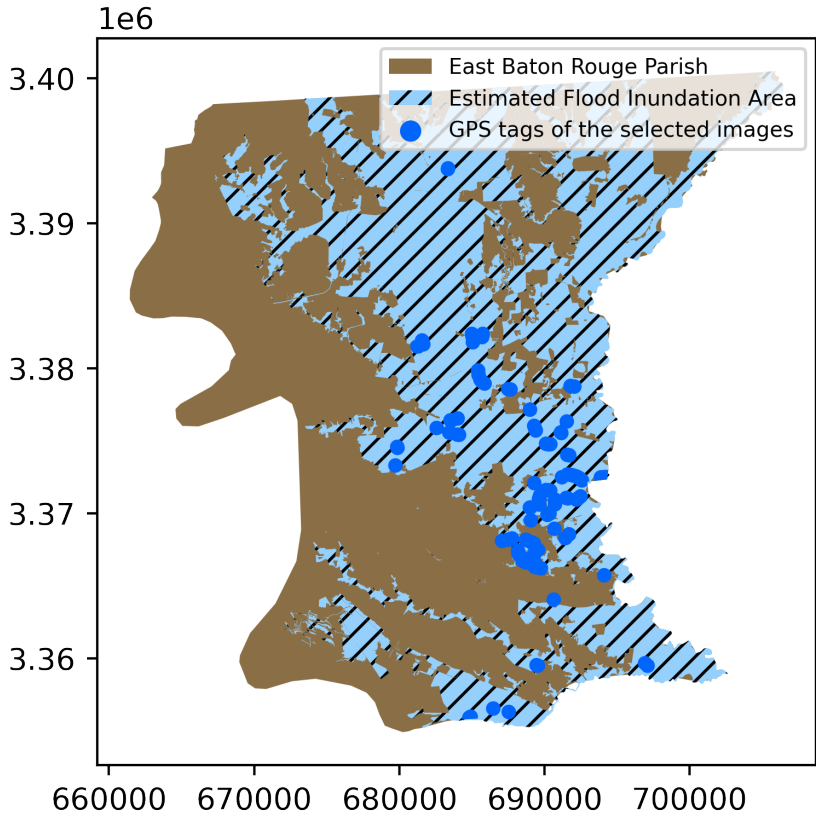


Figure 4-4: Flooding estimates and precision values using GPS tags (Precision: 52%)

of the flood images that lie in the FEMA estimates. Second, I estimate the flooding using the entire footprints of images classified as containing *"flooding/water."* Finally, I use my approach as the flood estimate. We only consider the flooding within the East Baton Rouge administrative boundary, since I do not have data on the flooding extent outside of the boundary.

Figure 4-4, Figure 4-5 and Figure 4-6 show the different flooding estimates overlaid against the official estimates, as well as the precision values of each method. These estimates were made with $\gamma_1 = 4$ and $\gamma_2 = 5$ km^2, so that ultimately 243 images were used. These results show a clear improvement going from using the GPS locations of the images to using the georeferenced footprint. This suggests that using the GPS tags of the images on their own is insufficient, since a large number of images containing flooding were taken over areas that were not flooded, and vice versa.

Furthermore, we see that my approach provides an improvement in precision compared to using the full image footprint. Figure 4-7 provides a close-up of the flooding

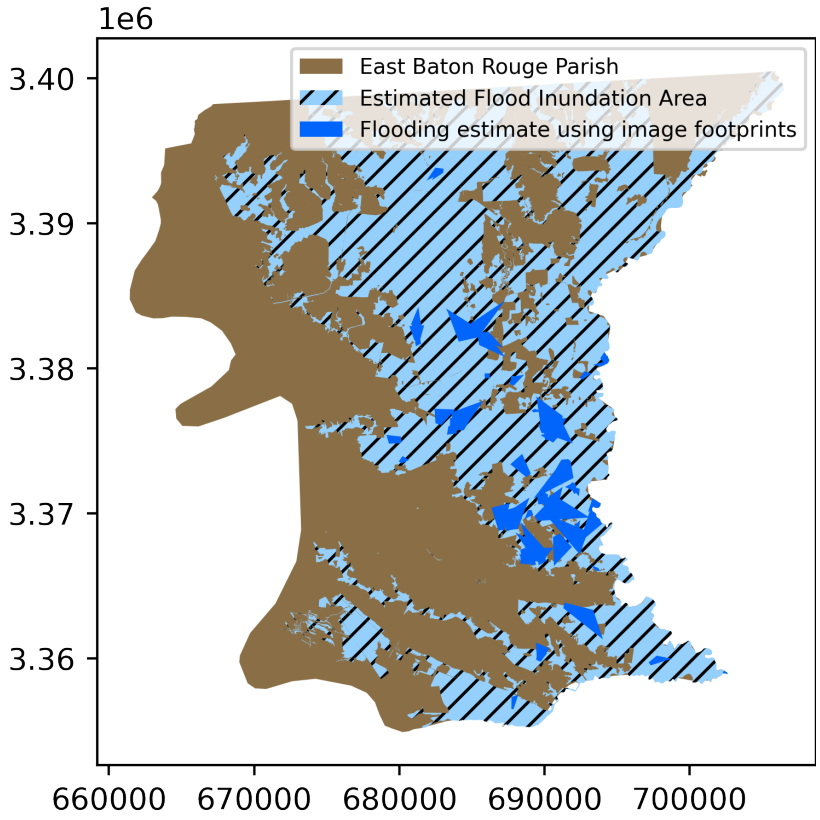


Figure 4-5: Flooding estimates and precision values using image footprints (Precision: 85%)

estimates for both methods, color-coded as true or false positive. Especially in areas at the edges of the flooding extent, my method provides a more precise outline of the official estimates than using the full image footprints.

To characterize the improvement that my approach provides over using the image footprints, I compute in Table 4.2 the difference in precisions $P_{\text{my approach}} - P_{\text{footprint}}$ for various values of γ_1 and γ_2, where P_m is precision of method m. We see that for all chosen combinations of thresholds, my approach has higher precision than the

		γ_2 (km^2)			
		1	**2.5**	**5**	**10**
γ_1 (unitless)	**2**	4	4	4	2
	3	5	4	2	2
	4	5	4	3	16
	5	5	4	9	19

Table 4.2: $P_{\text{my approach}} - P_{\text{footprint}}$ in percentage points for various values of γ_1 and γ_2.

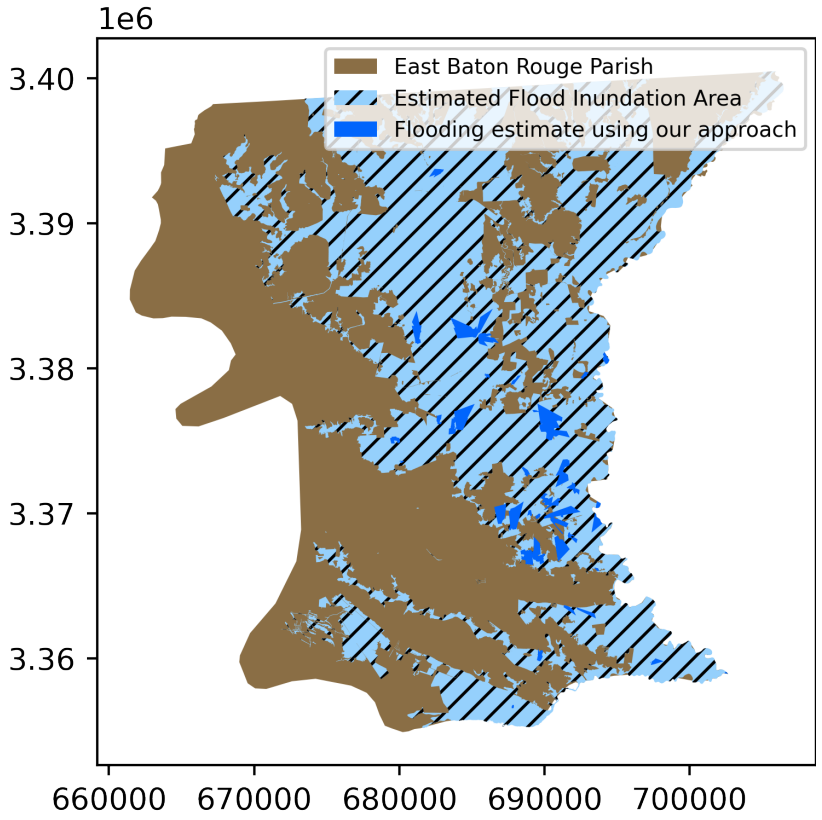
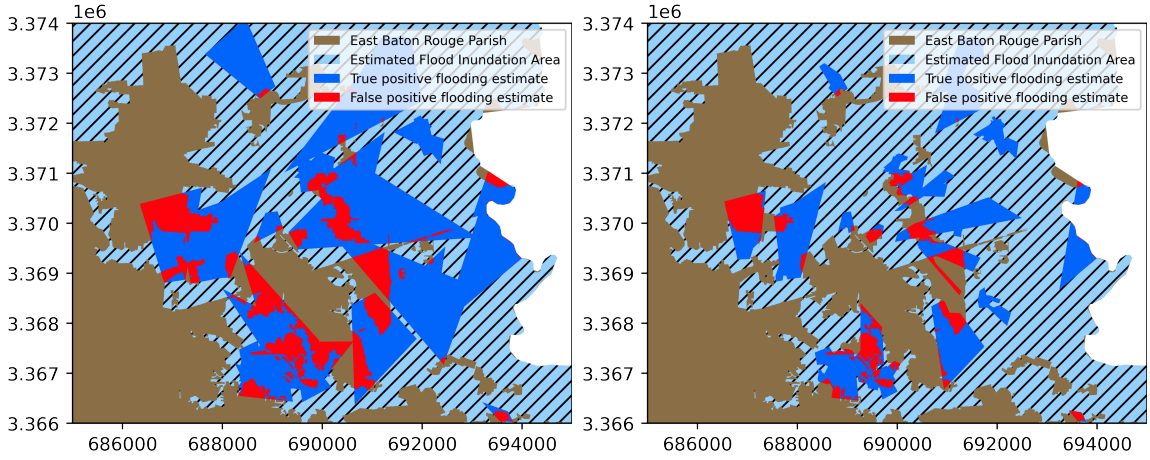


Figure 4-6: Flooding estimates and precision values using my approach (Precision: 88%)

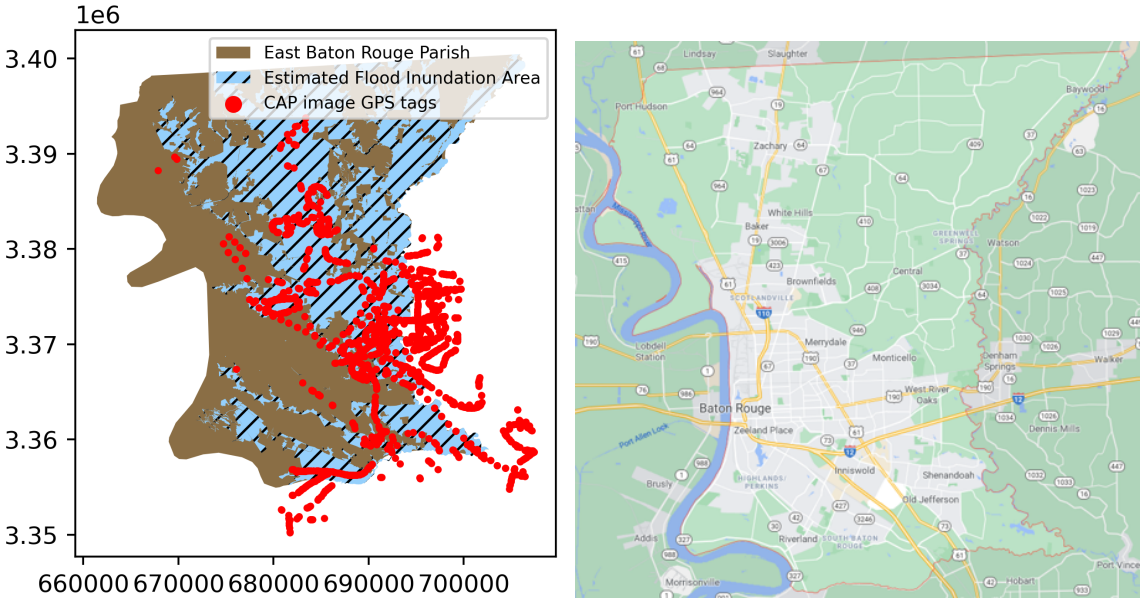
approach using only the image footprints. Furthermore, while the lowest prediction value I obtained with my approach was 86% (for $\gamma_1 = 5$ and $\gamma_2 = 5$), the lowest precision value of the image footprint was 68% ($\gamma_1 = 5$ and $\gamma_2 = 10$), suggesting that using the full image footprint is less robust to the choice of these parameters.

However, these maps also show that there are large portions of flooding that go undetected by either method. Indeed, my method only captures 3% of the total flooding in the scene. I posit that this may be a result of three key factors. First, there are some areas that are not detected by the image classifier as flooding. At worst the recall of my classifier is 79%, meaning that 21% of flooding images are not registered as flooding. Second, there could be large areas that are simply not imaged. Recall that CAP missions typically have a specific target set by a client. If the client does not include the entire parish, some areas will get left out. Figure 4-8 hints that this may be the case. Notice that while there are large rural areas that are flooded in



(a) Image footprint (b) My approach

Figure 4-7: Close-up of flooding estimates for both approaches, showing true and false positive regions.



(a) GPS tags (b) Google Maps screenshot of the parish

Figure 4-8: Comparison of GPS tags to a Google Maps screenshot of the parish.

the top right, these areas are barely imaged. Such blind spots might present equity concerns, as people that live in rural areas (or simply areas that are not seen as important by the client) may have to wait longer for CAP to detect that there is damage in their area. Finally, recall that it is not sufficient that an area is imaged. For georeferencing to work, two or more images have to be pointing at roughly the same scene. Therefore, if a CAP operator captures a single image of a scene or images that do not have enough overlap (which absent any guidelines is a reasonable action), those images would not be able to go through the entire DEL pipeline. Of the 1615 images, only 809 were reconstructed using structure from motion, meaning that roughly half of the images were thrown out to start with. Since images are further filtered when certain imaging criteria are met, this can lead to drastic reductions in the amount of usable imagery.

4.4 Moving past flooding

I chose to focus specifically on detecting flooding for two reasons. First, empirically I observed that it was the highest quality "Damage" label simply by observing whether an image labelled as flooding unambiguously contained flooding. Second, FEMA and others already rely on methods for detecting the extent of flooding after the fact, which made the quantitative evaluation of my approach much simpler. However, given that these methods exist, my approach is arguably more useful for "Damage" labels for which we do not have existing methods other than manual surveys, such as rubble.

While generalizing this to other labels is certainly possible, and I have designed my approach to be flexible to DEL of any kind of damage, I believe that currently this is operationally very challenging. Figure 4-9 shows the maximum, median and minimum precision performance of various classifiers when trained on the LADI dataset in TRECVID 2020, a popular video and image analysis and retrieval evaluation [4]. This plot shows that, while some labels such as *water* and *building* are easier to detect, others show very poor detection performance. Crucially, all "Damage" labels

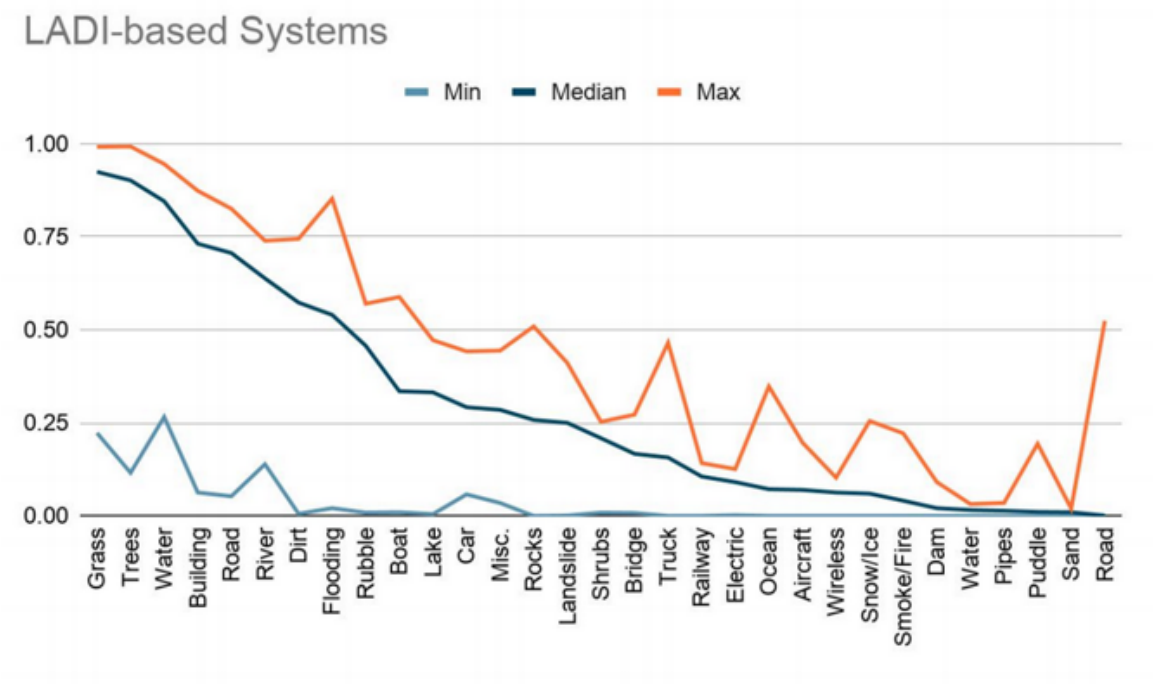


Figure 4-9: Maximum, median and minimum precision performance for various classifiers when trained on the LADI dataset in TRECVID 2020 [4].

except *flooding/water damage* have a median precision value less than 50%. Given that class activation mapping relies on classification performance, it is reasonable to believe that damage estimation would not work well with these labels. While it is difficult to know precisely why some labels are more easily recognizable than others, it is likely correlated with how often each label actually appears in the dataset (see Figure 2-1).

To conclude, I have shown that given a set of images from a disaster area, we can apply my approach to detect flooding with fairly high precision. However, there are still large portions of inundated areas that undetected by my approach. This is primarily due to poor imaging throughout the disaster region. Furthermore, a class activation mapping approach likely would not work for other types of damage such as rubble because of poor representation in the LADI dataset. All of this points to important work that needs to be carried out in order to fully operationalize CAP imagery. In the next and final chapter of this thesis, I will propose technical and policy interventions to address these shortcomings.

Chapter 5

Improving post-disaster imaging

In this thesis, I have jointly provided an examination of CAP's organizational characteristics as well as those of its post-disaster imaging program; and proposed an approach towards using their images to estimate and localize damage, a problem referred to as DEL. I showed that we can do this with high precision, but that there are still large portions of damage that could remain undetected. Furthermore, I have outlined some challenges regarding generalizing our approach towards additional damage labels, even though in principle it is possible to do.

This final chapter of my thesis will explore different technical and policy interventions that could improve the utility of the CAP imaging program. They are guided by the principle that we do currently have the capacity to quickly and effectively carry out DEL. Doing so is in line with both of the major motivations outlined in Chapter 1: saving human lives and increasing the value for money of taxpayer dollars.

5.1 Documenting best practices

In Chapter 2.1, we saw that CAP provides a comprehensive Aerial Photographer Task Guide that guides photographers on how to effectively image areas after disaster. Likewise, CAP offers full task guides for all other personnel involved in imaging missions. Of course, these guidelines are not suited for the imaging requirements laid out in Chapter 3.

The most immediate and straightforward intervention that can be put in place is to document best practices suitable for our DEL approach. This of course should happen in conversation with CAP volunteers to ensure that they are operationally feasible. Based on the imaging requirements listed in this thesis, these best practices would include:

1. Burst-imaging of important scenes to ensure they can be georeferenced
2. Use of the circular flight patterns shown in the Task Guide [31] to remove ambiguity in the 3D reconstruction (see Chapter 3.3.3)
3. Avoid taking images that are so oblique that they include the horizon
4. If possible, use cameras with orientation estimates (*e.g.* magnetic compasses) to aid in reconstruction
5. Seek opportunities to image areas that might be underserved (*e.g.* rural communities)

These best practices should be incorporated into CAP photographer training to ensure that all volunteers are aware of them. Furthermore, they should be periodically reviewed to ensure that new technological developments, such as the VIRB and WaldoAir imaging systems discussed in Chapter 2 are taken into account once they are more ubiquitous.

5.2 A multi-sensor approach

So far, we have only considered an approach that works solely off of CAP imagery to solve the DEL problem. It is worth pointing out that there is no reason to think that low-altitude aerial imagery alone is enough (or indeed will ever be enough) to detect all or even most instances of damage. Aerial imagery might get around occlusion from clouds, but there are still other forms of occlusion that still hamper the utility of aerial imagery, such as trees. Furthermore, it may be impossible to fully image

smoke and fire due to the flight hazard that smoke presents. Finally, even in the most equitable of flight plans it is essentially inevitable that some areas will lack priority and thus not be imaged.

That being said, modern post-disaster needs assessment does not rely on just one sensor. Rather, it takes into account input from different sensing systems to arrive at a more complete picture of the situation. For example, the East Baton Rouge flooding estimates came from a variety of sources, including 911 call-outs, Baton Rouge Fire Department search and rescue data, 311 requests for service, street-level damage assessments from City-Parish staff and other public officials, debris collection routes, road closure information, NOAA imagery, and FEMA DFIRM flood hazard areas [28]. There are other potential sources of data that could aid in needs assessment, including drone and satellite imagery.

CAP imagery has for the most part evaded use because, unlike many of the data streams described above, there was no automated way of interpreting the images being gathered. With the approach described in this thesis, CAP imagery could easily find its role within the larger needs assessment framework. Additional work needs to be conducted to fully understand what types of damage and under what circumstances CAP imagery contributes to post-disaster needs assessment relative to other types of imagery and data streams. Once this niche has been identified, I believe CAP imagery would provide an excellent value for money given how much information can be extracted and how inexpensive this process is.

5.3 CAP-to-satellite georeferencing

More than using multiple sensing systems in DEL, it may be even possible to incorporate other modalities into our own approach. For example, in Chapter 3.3, I outlined several conditions needed in order to georeference a CAP image. In summary, there need to be at least two images with enough overlap that common points between them can be automatically extracted. This actually puts a significant ceiling on how many images can be georeferenced. Absent any specific instructions to the contrary,

it is entirely reasonable that a volunteer would think that images that point at the same scene are a waste of effort. This is clearly visible from our case study, as almost half of the images (806 of out 1615) were unable to be georeferenced at all. Add the additional filters listed in Chapter 3.3.4 and you are throwing away a lot of potentially useful data. Because of this, it would be preferable not to have to rely on this method alone for georeferencing.

One approach that I have experimented with is georeferencing a CAP image against a satellite image. If possible, this would have very desirable properties. Satellite images already have a known, high-accuracy geotransform, which would allow us to directly georeference CAP image that we have found within the satellite tile. Furthermore, at this point the entire globe has been imaged via satellites, so finding a tile that corresponds to a CAP image is at least possible. However, one crucial limitation is that popular visual features such as ORB or SIFT have been shown to perform poorly under extreme changes in perspective [48, 37], as was the case when we attempted this. One could in theory establish ground truth correspondences between points on a CAP and satellite image and train some machine learning model on this dataset, but this would be extremely onerous and there are no obvious ways to automate this process.

I am currently working on a weakly supervised approach towards this that combines the general idea of class activation mapping [47] with the work on cross-view geolocalization of Workman et al and Liu et al [46, 21]. Cross-view geolocalization refers to the problem of finding the most likely place a ground-level image was taken given a series of satellite images. Note that this problem is distinct from that of georeferencing, which seeks to estimate a geotransform for a given image, and also note that this problem is irrelevant to our specific application given that we already have coarse GPS estimates for our images. This problem has typically been solved by the authors and by others by using Siamese neural networks to establish similarity between ground-level and satellite images.

If a neural network model can give a measure of similarity, it is also possible that it can distinguish what specific aspects of a pair of images are similar. For example,

it could point to a tree or a house in both images. This is the same idea that has been used in class activation mapping for detecting which areas of an image correspond to a specific class [47]. I believe it is possible to exploit this property that has so far been used mostly for image classification to detect similar patches between images. In the specific case of a CAP and satellite pair, knowing what areas of an image correspond to another can yield an estimate of a geotransform. While this idea has not been fully tested, I believe that this approach could help lift the current ceiling preventing us from using a greater number of CAP images in our damage estimates.

5.4 Investing in CAP imaging infrastructure

My final recommendation is more aspirational and revolves around investment. CAP imagery is rarely used in disaster response and there are little signs of government interest in investing effort and funding towards improving the program. Given that it is by far the cheapest form of aerial surveillance after a disaster and that we have shown that it is indeed possible to perform DEL using solely CAP imagery, I believe the program is worth further government investment.

There are two main areas I believe would benefit from additional government interest. The first is imaging equipment. As discussed in Chapter 2, all aircraft are equipped with handheld DSLR cameras which is what is currently being used for disaster imaging. A much, much smaller subset of cameras are outfitted with VIRB or WaldoAir systems, which are currently being tested to see if wider adoption is warranted [15]. There is little reason to believe that these would not be a large improvement over the status quo. However, the Department of Homeland Security has indicated that it has no intention in the short term of phasing out the current DSLR cameras in favor of the more useful VIRB and WaldoAir systems.

I believe that this is a mistake. Completely eliminating the georeferencing ceiling would immensely improve the utility of CAP imagery and therefore lead to a greater amount of damage detected in a short amount of time. While we do not make any economic analyses on the tradeoff between the cost of these sensors and the potential

safeguarding of life and property, I believe it is appropriate for CAP to at least have one aircraft equipped with either one of these sensors per state, with a potential for more in states that are more often affected by natural disasters. This would ensure that CAP retains the capability to conduct proper surveillance after these events.

The second area worth investing on is low altitude imaging datasets. With the exception of LADI [20], there are no publicly available datasets for low-altitude disaster imaging. On the contrary, there are no shortage of satellite and orthorectified aerial imagery datasets for a variety of purposes, including disaster response. Given our results, I believe it would be appropriate to invest in more complete, accurate, and robust datasets. In particular, I believe that there should exist datasets that are completely labelled by experts in the disaster response community and that include bounding boxes or segmented images. In my conversations with the Humanitarian Assistance and Disaster Relief Group and MIT Lincoln Laboratory I have been made aware of efforts to augment LADI with bounding boxes. However, funding for the project has not materialized. I believe investing in these improvements to low-altitude disaster datasets would improve the accuracy and performance of our damage estimation pipeline.

5.5 Concluding remarks

I began this thesis talking about the steps in FEMA's Recovery Continuum. Specifically, I mentioned that the steps within the Continuum are interrelated, and that it would be unwise to consider one without consideration for the others. Many of the steps outlined in this final chapter are preparedness steps, and it is of crucial importance that we take these steps as soon as possible since more effective short-term recovery would indeed ensure a more effective medium- and long-term recovery as well. This involves a holistic approach involving multiple stakeholders, areas of expertise and technologies, and a crucial component of that is CAP.

I believe that CAP imagery has fallen into something of a vicious cycle. Because it has not up until this point been a reliable stream of information for post-disaster

needs assessment, it has received little interest from the disaster response community and the federal government. As a result, the program cannot improve which in turn decreases its utility, thus perpetuating the cycle.

In working towards this thesis, my goal has been to break this cycle; my goal has been to show that CAP imagery can indeed serve as a vital tool for aerial surveillance following a natural disaster. I have shown that even in its current state, we can estimate and localize damage with very high precision. My expectation now is that this work can be used as a tool for advocacy on various fronts. For CAP, this thesis can advocate for new processes that would lend themselves better towards DEL. For the federal government, this thesis can advocate for renewed interest and funding to improve this important program. And for academia, this thesis shows that there is exciting and important work to be done in disaster response. All of these fronts are of crucial importance, especially in the face of increasing frequency and intensity of natural disasters. In the end, renewed interest in this line of work can help save lives.

Appendix A

Additional Disaster Imagery Datasets

1. *xBD*: A satellite imagery disaster dataset. It shows areas before and after a natural disaster. Contains building footprints, level of damage for each building, bounding boxes for different indicators of damage, and image metadata (including geotransform) [12].

2. *FloodNet*: A UAV image flooding dataset, taken after Hurricane Harvey. Shows segmented images for a variety of pixel-level classes, including building, flooding, and trees [34].

3. Benchmark Dataset from Chen et al: A dataset containing aerial images from the U.S. National Oceanic and Atmospheric Administration and satellite images from DigitalGlobe taken after Hurricane Harvey. Contains the raster data as well as vector data with building damage estimates [6].

4. GEO-CAN Christchurch dataset: A dataset containing post-earthquake data from the Christchurch earthquake in 2011. It includes aerial imagery as well as building damage bounding boxes that were gathered by an on-the-ground survey team [10].

5. *Volan2018*: An aerial disaster imagery dataset that contains bounding box annotations for various on-the-ground objects, including debris and flooding. Other than LADI, this is the only dataset that I am aware of that contains

oblique imagery. As of the writing of this thesis, this dataset is not freely available [33]

Bibliography

[1] *36 U.S. Code Chapter 403 - CIVIL AIR PATROL.*

[2] *Land Remote Sensing Policy Act of 1992 (1992 - H.R. 6133)1992.* 1992.

[3] Alex M Andrew. *Multiple view geometry in computer vision.* Emerald Group Publishing Limited, 2001.

[4] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Luca Rossetto, Heiko Schuldt, George Awad, Asad A Butt, George Awad, et al. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks. *Proceedings of TRECVID 2020*, 32:14, 2020.

[5] J-L Bessis, Jerome Bequignon, and Ahmed Mahmood. The international charter "space and major disasters" initiative. *Acta Astronautica*, 54(3):183–190, 2004.

[6] Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] Mark A Fonstad, James T Dietrich, Brittany C Courville, Jennifer L Jensen, and Patrice E Carbonneau. Topographic structure from motion: a new development in photogrammetric measurement. *Earth surface processes and Landforms*, 38(4):421–430, 2013.

[10] R Foulser-Piggott, R Spence, K Saito, DM Brown, and R Eguchi. The use of remote sensing for post-earthquake damage assessment: lessons from recent events, and future prospects. In *Proceedings of the Fifthteenth World Conference on Earthquake Engineering*, page 10, 2012.

[11] Hernâni Goncalves, Luís Corte-Real, and José A. Goncalves. Automatic image registration through image segmentation and SIFT. *IEEE Transactions on Geoscience and Remote Sensing*, 49(7):2589–2600, Jul 2011.

[12] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

[13] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Desmarais John and David Reichert. 2D & 3D rapid collection imagery testing, Oct 2020.

[16] Christopher Joyce. New report shows weather disasters in 2017 cost more than $300 billion. *NPR.org*, Jan 2018.

[17] Dong-Ki Kim and Matthew R Walter. Satellite image-based localization via learned embeddings. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2073–2080. IEEE, 2017.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[19] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[20] J. Liu, D. Strohschein, S. Samsi, and A. Weinert. Large scale organization and inference of an imagery dataset for public safety. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6, Sep. 2019.

[21] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5633, 2019.

[22] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[23] mapillary. OpenSfM. https://github.com/mapillary/OpenSfM, Feb 2021.

[24] David Martin. Private company launches "largest fleet of satellites in human history" to photograph earth. *CBS News*, Jan 2019.

[25] Theo Moons, Luc Van Gool, and Maarten Vergauwen. *3D reconstruction from multiple images: principles*. Now Publishers Inc, 2009.

[26] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009.

[27] Air Forces Northern. Civil air patrol completes full year of covid-19 support. *U.S. Air Force*, Mar 2021.

[28] City of Baton Rouge and Parish of East Baton Rouge. The great flood of 2016 story map, Aug 2016.

[29] United States Department of Homeland Security. *National Disaster Recovery Framework, Second Edition*. Jun 2016.

[30] Jaehong Oh, Charles K Toth, and Dorota A Grejner-Brzezinska. Automatic georeferencing of aerial images using stereo high-resolution satellite images. *Photogrammetric Engineering & Remote Sensing*, 77(11):1157–1168, 2011.

[31] Civil Air Patrol. *Airborne Photographer Task Guide*. Civil Air Patrol, May 2013.

[32] Civil Air Patrol. *CAPabilities Handbook: A Field Operations Resource Guide*. Civil Air Patrol, Nov 2013.

[33] Yalong Pi, Nipun D. Nath, and Amir H. Behzadan. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43:101009, January 2020.

[34] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Murphy. FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *arXiv preprint arXiv:2012.02951*, 2020.

[35] Thomas Reilly. Florida's flying minute men: The Civil Air Patrol, 1941-1943. *The Florida Historical Quarterly*, 76(4):417–438, 1998.

[36] Frances Robles, Kenan Davis, Sheri Fink, and Sarah Almukhtar. Official Toll in Puerto Rico: 64. Actual Deaths May Be 1,052. *The New York Times*, Dec 2017.

[37] Akshay Shetty and Grace Xingxin Gao. UAV pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1827–1833. IEEE, 2019.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985.

[40] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017.

[41] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[42] CRED UNISDR et al. The human cost of natural disasters: An overview of the last 20 years. 2019.

[43] Jana Wäldchen, Michael Rzanny, Marco Seeland, and Patrick Mäder. Automated plant species identification—trends and future directions. *PLoS computational biology*, 14(4):e1005993, 2018.

[44] Chun-Po Wang, Kyle Wilson, and Noah Snavely. Accurate georegistration of point clouds using geographic data. In *2013 International Conference on 3D Vision-3DV 2013*, pages 33–40. IEEE, 2013.

[45] Debra Werner. Satellites to the rescue after natural disasters. *SpaceNews*, Jun 2019.

[46] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[48] Xiangyu Zhuo, Tobias Koch, Franz Kurz, Friedrich Fraundorfer, and Peter Reinartz. Automatic UAV image geo-registration by matching UAV images to georeferenced image data. *Remote Sensing*, 9(4):376, 2017.