THE COLLEGE OF WILLIAM & MARY

DISSERTATION PROSPECTUS

Studying Global Conflict with Deep Learning and Satellite Imagery

---

Author:
Scott WARNKE

Advisor:
Dan RUNFOLA

A prospectus submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the

Data Science
Department of Applied Science

April 2, 2024

Contents

List of Figures

List of Tables

List of Abbreviations

ACLED	Armed Conflict Location & Event Data Project
CAM	Class Activation Maps
CIFAR-10	Canadian Institute For Advanced Research - 10 classes
CIFAR-100	Canadian Institute For Advanced Research - 100 classes
CNN	Convolutional Neural Network
DEGURB	Degree of Urbanisation
GDP	Gross Domestic Product
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
LLM	Large Language Model
NIR	Near-Infrared
OBIA	Object Based Image Analysis
RGB	Red, Green, and Blue
ResNet	Residual Neural Network
SAT-CAM	Satellite Class Activation Mapping
Score-CAM	Score-weighted Class Activation Mapping
SSGrad-CAM	Spatially Sensitive Gradient-weighted Class Activation Mapping
TB	Terabyte
U.N.	United Nations
VGG	Visual Geometry Group

Chapter 1

Introduction

Instances of social unrest, often manifesting as riots or protests, wield significant influence on the communities, regions, and nations in which they unfold (Bencsik, 2018). The repercussions of such events are wide-ranging, ranging from geopolitical transformations (i.e., riots in Egypt in 2011 (Joya, 2011) and Hong Kong in 2019 (Purbrick, 2019)) to substantial economic losses (exemplified by the hundreds of millions of dollars incurred during the 2011 riots in the UK (Bencsik, 2018)). These events may result in human casualties, as evidenced by food riots in Africa in 2007-08 (Berazneva and Lee, 2013) and riots caused by garbage collection issues in Beirut in 2015 (El Warea et al., 2019). These events impact cities across the entire globe, with recent examples in Latin America (Eckstein, 2001), Asia (Purbrick, 2019), Africa (Joya, 2011; Berazneva and Lee, 2013), and Europe (Andronikidou and Kovras, 2012). Because of the importance of these events, scholars across multiple disciplines have sought to both predict and understand them, using a wide range of data sources and techniques (Pond and Lewis, 2019; Snow, Vliegenthart, and Corrigall-Brown, 2007; Davies et al., 2013). However, most of these approaches have relied on sources that may not be available or reliable in geographies of interest, such as news articles. Here, we explore the capability of satellite imagery to aid in the prediction of protest and riot events, explicitly seeking to understand the degree to which this globally-available source of information may be able to augment existing predictive methodologies. This approach exploits correlations between the human built environment - i.e., urban form (Fox and Bell, 2016) - and the likelihood of a protest or conflict event at a given geographic location.

One of the core innovations that enables us to estimate social events (such

as conflict) using satellite imagery is convolutional modeling (Goodman, BenYishay, and Runfola, 2021). Deep learning, including the use of Convolutional Neural Networks (CNNs), is being used in a wide range of applications from estimating school test scores (Runfola, Stefanidis, and Baier, 2022) and predicting poverty rates (Jean et al., 2016), to detecting changes in urban environments (Daudt et al., 2018) and tracking typhoons (Rüttgers et al., 2019). This includes innovations from the field of computer vision, which have shown the capability of CNNs to detect objects (Shin et al., 2016), classify images (Krizhevsky, Sutskever, and Hinton, 2017), and recognize images (Chauhan, Ghanshala, and Joshi, 2018). Deep learning can be used in conjunction with satellite imagery to perform many different classification and detection tasks, such as detecting infrastructure destruction in conflict environments (Nabiee et al., 2022), identifying ships (Leclerc et al., 2018; Patel, Bhatt, and Mazzeo, 2022), land cover and land usage analysis (Helber et al., 2019; Kussul et al., 2017; Carranza-García, García-Gutiérrez, and Riquelme, 2019; Lv et al., 2024), urban expansion (Zhang et al., 2018; He et al., 2019a; Zhang et al., 2019), and road quality analysis (Brewer et al., 2021). Building on this work, in this piece we combine global-scope high resolution satellite imagery sourced from *Planet* with information on the spatial distribution of protest and riot events from *ACLED*, seeking to establish the degree to which satellite information can be used to directly predict the geospatial locations of protest events.

In addition to this core aim, we further seek to advance our ability to understand the specific elements of urban form that the model finds are correlated with conflict. This requires the development of new tools and techniques - broadly referred to as explainability techniques (Buhrmester, Münch, and Arens, 2021) - that are interoperable with satellite imagery. Today, state of the art techniques such as Class Activation Maps (CAM) (Zhou et al., 2016) or Deconvolution/inverting neural networks (Noh, Hong, and Han, 2015) are under-developed in the context of satellite imagery (Vasu, Rahman, and Savakis, 2018; Charuchinda et al., 2019; Fu et al., 2019). Paper 2 will focus on overcoming these limitations.

An important aspect of utilizing satellite imagery in conjunction with deep learning, is the ability of techniques to scale. If we are able to predict conflict, but only on small scales that compare small standardized images, our applications are limited. We aspire to identify *where* conflict will occur from full satellite

scenes that cover hundreds of square kilometers. Simply stated, are we able to train a neural network to identify the localized location of conflict from a full satellite scene. This will necessitate new training methodologies to appropriately segment large satellite images into smaller portions, in an attempt to identify potential hot spots for future conflicts. These new methodologies will come with associated challenges that relate to training neural networks to handle segmented and subsetted data from full satellite images, evaluating segmented images as part of larger satellite scenes to determine where hot spots exists, and potentially the requirement to generate more data for training and testing.

With these three papers, we will explore the question *does satellite imagery contain information that can be used to estimate the likelihood of conflict across geographic locations?* To explore this question three sub-questions will be explored:

Q1. *Can satellite imagery alone be used to determine the likelihood of conflict in urban areas?*

Q2. *Can we semantically describe the features within a satellite image that are consequential to the classification of urban conflict?*

Q3. *Can deep learning techniques enable the localization of conflict across a full-sized satellite image?*

In this prospectus, I provide an initial quantitative analysis focused around **a convolutional neural network's ability to identify conflict**, designed to illustrate my quantitative capabilities as well as preliminary results. For the following two research questions, I provide theory, datasets, and methods that I propose to test. The prospectus is structured as follows. In chapter 2, I introduce literature common to each of these research topics. In chapter 3, I introduce my first research question, inclusive of a preliminary quantitative analysis. In chapter 4, I introduce my future plans for research questions 2 and 3. Finally, in chapter 5, I provide an outline of the anticipated schedule for graduation, as well as description of risks that may inhibit my progress and efforts to mitigate those risks.

Chapter 2

Literature Review

2.1 Measurement and Modeling of Riots and Protests

Riots and protests constitute integral components of democratic societies (Anderson and Mendes, 2006), yet it is imperative for government authorities to effectively mitigate the economic and human costs that may be associated with these events to maintain stable governance (Klein and Regan, 2018). This is accentuated by the heightened prevalence of protests and riots on a global scale in recent years (Ciorciari and Weiss, 2016). One viable strategy for authorities to temper the negative impacts of these events is through preemptive allocation of resources, such as medical units (Gong and Batta, 2007) or increased international presence (i.e., U.N. peacekeepers) in anticipation of unrest (Greer and McLaughlin, 2010). At the international scale, in an attempt to protect citizens who are traveling abroad, responsible governmental foreign offices (the US Department of State as an example) may also issue travel warnings for particular areas to avoid (Löwenheim, 2007). However, proactive approaches necessitate the capacity to predict both the time and location of potential conflict events (Wu and Gerber, 2017).

A number of approaches exist which aid in the measurement and prediction of protests or riots (Wu and Gerber, 2017). Past literature, for instance, has demonstrated the utility of news reports in providing valuable insights into civil conflict, such as riots and protests in response to rising food prices (Heslin, 2021). Using this approach, studying riots in France, researchers were able to replicate the spread of riots using an epidemic-like model with as few as six parameters that included population demographics, police reports, and spatial

information (Bonnasse-Gahot et al., 2018). Social media platforms represent another venue for authorities to detect and analyze real-world events, including social unrest like riots and protests (Becker, Naaman, and Gravano, 2011; Korolov et al., 2016; Petrović, Osborne, and Lavrenko, 2010). X (formerly Twitter) is a common focus of these studies, and can be used as a near real-time reporting source (Becker, Naaman, and Gravano, 2011). Analysis of twitter data has demonstrated the correlative relationship between daily hashtag use and protests, predicting protests 24-48 hours prior to occurring in Baltimore and New York City during 2015 (Korolov et al., 2016). Prior work in this field has show the ability to predict the probability of fatalities associated with conflict events using satellite imagery, within conflict areas in Nigeria, with accuracy rates of 80% when combining Landsat imagery and CNNs (Goodman, BenYishay, and Runfola, 2021).

Much of the current research in forecasting social unrest is focused on the likelihood of a future event (Renaud et al., 2019; Phillips et al., 2017; Cadena et al., 2015; Filchenkov, Azarov, and Abramov, 2014; Compton et al., 2013). There are other efforts to better understand and model the characteristics of smaller sub-events within broader riots, such as shooting or fires (Alsaedi, Burnap, and Rana, 2017). Modeling of riots demonstrates an ability to accurately simulate many of the spatial characteristics of riots, including the distance participants will travel within contiguous riot areas (Davies et al., 2013). Beyond riot forecasting, tweet analysis demonstrates the ability to detect and discriminate between disruptive events within a riot and normal information dissemination (Alsaedi, Burnap, and Rana, 2015). Additionally, when analyzing individual behavior, social media has been studied to demonstrate not only how information is distributed about future and concurrent protests, but also how individuals are recruited into protesting (González-Bailón et al., 2011).

The accuracy and spatial specificity of existing riot and protest forecasting techniques vary. Previous research has shown that leveraging information from social media (i.e., Tweets) can result in the accurate prediction of riots in some cities (i.e., Baltimore and New York City), but these models require location-specific information or hashtags which inhibit their use in other settings (i.e., San Francisco) (Korolov et al., 2016). Related tweet-based analyses have shown that accurate temporal estimates across broad geographies are possible, but without

spatial specificity in where riots or protests are likely to occur (González-Bailón et al., 2011). Other researchers have used a broader range of sources to achieve higher spatio-temporal accuracy, such as police reports, but these techniques are inherently limited to a small number of areas where such information is available (Alsaedi, Burnap, and Rana, 2015; Korolov et al., 2016; González-Bailón et al., 2011; Bonnasse-Gahot et al., 2018; Alsaedi, Burnap, and Rana, 2017).

2.2 Satellite Imagery

There is a long history of utilizing satellite imagery in research that is based on visually observable characteristics, such as habitat and land cover change (Alo and Pontius Jr, 2008; Stow et al., 2008; Rogan and Chen, 2004), soil evaluation (Foody and Mathur, 2004), and urban land cover (Zhou and Troy, 2008). When satellite imagery is used in conjunction with deep learning techniques, including CNNs, researchers have recently been able to learn about topics not normally associated with traditional satellite imagery uses, such as predicting crime (Najjar, Kaneko, and Miyanaga, 2018) or the prevalence of cancer (Bibault et al., 2020). Other examples include estimating human migratory flows (Runfola et al., 2022), estimating educational outcomes (Runfola, Stefanidis, and Baier, 2022), tracking economic growth in China (Brewer, Lv, and Runfola, 2023), predicting road quality (Brewer et al., 2021), and estimating socioeconomic census variables from satellite imagery (Runfola et al., 2024).

The ability to estimate income with satellite imagery is one of the most commonly studied topics, with researchers initially exploring the use of nighttime lights as a proxy for development (Elvidge et al., 2012). Building on these approaches, researchers leveraged a CNN in conjunction with both daytime satellite imagery and nighttime lights, an approach which was able to explain up to 75% of the of the variation of household wealth in various countries that lacked detailed survey data (Jean et al., 2016). A similar CNN-based approach was able to leverage satellite imagery to predict GDP and total retail sales with a Pearson coefficient of 0.85 for both tasks, outperforming linear regression models (Wu and Tan, 2019). CNNs that were trained on higher resolution imagery were able to explain 57% of the variation in poverty of municipalities in Mexico (Babenko et al., 2017). These authors used two different satellite image sources, MAXAR

(formerly Digital Globe) and Planet, noting a slight decrease in performance of Planet imagery but highlighting the benefit of Planet's daily image capability. When using publicly available satellite imagery with CNNs, researchers were also able to predict school test scores with accuracy's between 76% to 80% with images of the schools (Runfola, Stefanidis, and Baier, 2022). Even human migratory flows have been predicted with an accuracy of $R^2 = 0.72$ when satellite imagery and and census data are used in conjunction with CNNs, which outperforms the use of socioeconomic census data alone (Runfola et al., 2022). Some of the most recent literature on this topic has explored the ability of CNN based satellite models to estimate broader ranges of census variables in regions that may lack regular census instruments (Runfola et al., 2024).

In scenarios where data is challenging or impossible (i.e., historic time periods) to collect, there is increasing evidence that satellite imagery can aid in filling data gaps (Goodman, BenYishay, and Runfola, 2021; Jean et al., 2016; Bharti and Tatem, 2018; Hu et al., 2019; Aung et al., 2021). This characteristics of satellite information is important in the context of studying riots and protests, as the majority of literature we've identified has focused on the use of news or social media sources (Purbrick, 2019; Greer and McLaughlin, 2010; Wu and Gerber, 2017; Ciorciari and Weiss, 2016; Becker, Naaman, and Gravano, 2011; Korolov et al., 2016; Renaud et al., 2019; Phillips et al., 2017; Cadena et al., 2015; Filchenkov, Azarov, and Abramov, 2014; Compton et al., 2013; Alsaedi, Burnap, and Rana, 2017). There are many countries of research interest that do not allow free access to social media or control the news narrative, such as Russia (Gehlbach, 2010), China (Tai, 2014), Iran (Rahimi, 2015), and Venezuela (Pain and Korin, 2021). Satellite imagery provides a unique capability to access data in a country that might restrict access to social media or control news sources, motivating us to use satellite imagery to predict conflict.

There a many options to consider when using satellite imagery, (Kramer and Cracknell, 2008). A comprehensive evaluation of which imagery product to use broadly includes a consideration of it's spatial and spectral resolution, temporal revisit times, and radiometric precision (Xie, Sha, and Yu, 2008). A number of satellites and sensors have been launched over the last four decades, with many of these still providing regular imagery today. A list of potential sources for imagery is shown in table 2.1

Source	Number of Bands	Resolution	Revisit Time
Landsat 9	9 spectral bands	30m	16 Days
Sentinel-2	4 bands	10m	5 Days
SPOT-6	4 bands	6m	26 days
PlanetScope	8 spectral bands	3-4m	Daily
GeoEye-1	4 bands	1.84m	2-3 Days
WorldView-3	8 spectral bands	0.31m	1-4 Days

TABLE 2.1: Landsat data is for Landsat 9 (USGS, 2022–); Planet data is for PlanetScope (Planet, 2022b); WorldView-3 data can be found at (Maxar, 2020b); GeoEye-1 data can be found at (Maxar, 2020a); Sentinel-2 data can be found at (European Space Agency, 2015a); SPOT-5 data can be found at (European Space Agency, 2015b)

In this dissertation, we focus on the use of PlanetScope, a product provided by the commercial provider *Planet*. PlanetScope is a constellation of approximately 130 satellites, cumulatively capable of taking approximately 3-4 meter pixel resolution images across the landmass of the Earth each day (Planet, 2022a). To date there have been three generations of PlanetScope satellites, that have images dating back to July 2014. All three generations of the satellites have produced 3 band imagery (red, green, blue, channels), with the newest generation collecting 8 bands (Planet, 2022a). Each image covers between 25 x 11.5 km to 32.5 x 19.6 km, depending on the sensor used (Planet, 2022a; see table 2.2 for more details). In order to leverage the full length of time covered in our record of conflict events (see Chapter 3), we will use only the Red, Green, and Blue bands of data that have been provided by the earliest Dove Classic, Dove-R, and newest SuperDove satellites. An example of an image from Planet is shown in figure 2.1.

2.3 Convolutional Modeling

Computer vision is a branch of machine learning that attempts to train computers to visually learn and identify objects similar to vision in humans (Goodfellow, Bengio, and Courville, 2016). Early attempts in this field were pioneered by Rosenblatt using perceptrons to attempt to identify letters visually (Rosenblatt, 1957). These early concepts, like perceptron, would develop and advance

FIGURE 2.1: Satellite Image of Athens Greece, taken 31 January 2018. Imagery ©Planet Labs PBC 2023. All rights reserved.

	Dove Classic	**Dove-R**	**SuperDove**
Band	Wavelength (nm)	Wavelength (nm)	Wavelength (nm)
Red	590 - 670	650 - 682	650 - 680
Green	500 - 590	547 - 585	547 - 583
Blue	455 - 515	464 - 517	465 - 515
Image Area	25 x 11.5 sq km	25 x 23 sq km	32.5 x 19.6 sq km
Availability	July 2014 - April 2022	March 2019 - April 2022	March 2020 - present

TABLE 2.2: Technical Data from Planet about Planetscope(Planet, 2022b). As the Planetscope satellites have increased their capabilities over time, the technical specifications have slightly shifted. Later generations of the satellites have more bands available, but to maintain continuity over our data set, we are only considering RGB bands.

towards modern applications and the various forms of Deep Neural Networks (Fradkov, 2020). These neural networks were not limited to a single perceptron or neuron, but could have many layers with complex connections among the layers. Computational power and speed has improved significantly since the 1950s (Nordhaus, 2001), and so has the advancement of neural networks.

In this study, we rely on convolutional neural networks (CNN), a type of deep learning that is designed for the analysis of image data. These techniques have been shown to be effective at detecting, labeling and differentiating objects (Simonyan and Zisserman, 2014; Zhang, Zhang, and Du, 2016; He et al., 2016a; Krizhevsky, Sutskever, and Hinton, 2017; Voulodimos et al., 2018). CNNs represent a family of deep learning techniques that implement convolutional layers that extract features from an image (Zhang, Zhang, and Du, 2016). There are many types of CNN architectures that have performed well across a wide range of computer vision tasks (Voulodimos et al., 2018; Simonyan and Zisserman, 2014; Szegedy et al., 2015). In general, convolutional networks refer to functions that take image inputs and kernels to generate outputs known as feature maps (Goodfellow, Bengio, and Courville, 2016). A formal definition is shown in formula 2.1. The feature map, $S(i,j)$, is the output of image I with dimensions i,j, convolved with kernel K with dimensions m,n (Goodfellow, Bengio, and Courville, 2016).

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{2.1}$$

An early CNN that leveraged GPUs was AlexNet (Krizhevsky, Sutskever, and Hinton, 2017). The architecture of AlexNet comprises five convolutional layers and three fully connected layers. Despite its relatively shallow structure, this network exhibits exceptional performance and played a crucial role in promoting the adoption of CNNs in computer vision. Notably, advancements in AlexNet demonstrated that the integration of dense convolutional layers could enhance performance (Szegedy et al., 2015). A diagram of the architecture of AlexNet is displayed in figure 2.2.

Building upon this progress, a novel CNN named VGG was introduced by the vision geometry group at the University of Oxford (Simonyan and Zisserman, 2014). The VGG architecture emphasizes stacked convolutional layers

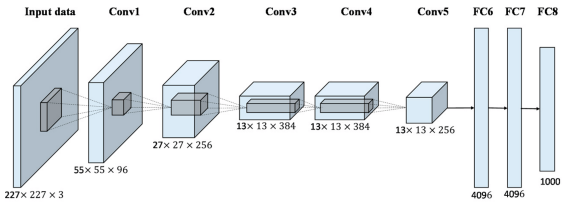


FIGURE 2.2: Architecture of AlexNet, highlighting input data passing through five convolutional layers and 3 fully connected layers. Figure is from (Han et al., 2017).

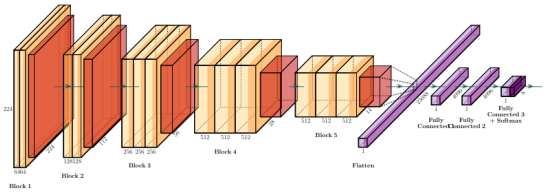


FIGURE 2.3: Architecture of VGG, highlighting input data passing through stacked convolutional layers before flattening and passing through fully connected layers. Figure is from (Vrbancic, Zorman, and Podgorelec, 2019).

with smaller kernel sizes, departing from individual convolutional layers with larger kernels. This approach yields significant computational gains and enables a more discriminative decision function through heightened non-linearity. Subsequent modifications and refinements to the VGG implementation have further improved the architecture's performance (Hu, Shen, and Sun, 2018). A diagram of VGG is shown in figure 2.3.

As VGG was developed, a desire to test so-called 'deeper' networks took hold as a potential route forward for improved levels of capability (Simonyan and Zisserman, 2014). However, VGG, AlexNet, and other earlier networks were functionally limited in the depth they could achieve due to - at the time - the difficulty of optimizing deep networks due to vanishing gradients (He et al., 2016a). ResNet, short for residual convolutional neural network, was developed to address the challenges associated with deeper networks while aiming to improve accuracy. A notable characteristic of ResNets is the inclusion of residual connections, which establish a direct path from one layer to deeper layers, bypassing intermediate convolutions (see figure 2.4). By incorporating

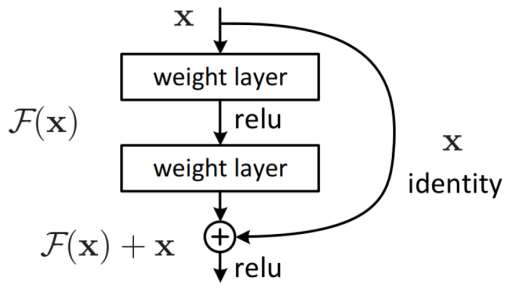


FIGURE 2.4: Residual learning block, highlighting information flow deeper into networks that by passes itermediate layers. Figure is from (He et al., 2016a).

these residual connections, the network can increase in depth without encountering issues related to gradient back propagation during optimization (He et al., 2016a). Following the initial introduction of ResNets by He et al., subsequent researchers have made various improvements and modifications, often altering specific dimensions while preserving the general ResNet structure (Xie et al., 2017; Zagoruyko and Komodakis, 2016). Furthermore, others have explored alternative approaches to enhance the training process or optimize the loss functions employed in ResNets (He et al., 2016b; Huang et al., 2016; He et al., 2019b; Wightman, Touvron, and Jégou, 2021). The concept of forwarding information through skip connections has also been expanded into fully connected dense blocks in the DenseNet architecture (Huang et al., 2017). Due to their remarkable effectiveness, ResNets have emerged as one of the dominant CNN architectures widely employed in contemporary deep learning applications.

By 2016, ResNet architectures reached depths of 1001 layers, and achieved error percentages as low as 4.62% on CIFAR-10 and 22.71% on CIFAR-100 (Xie et al., 2017). Research has demonstrated ResNet-18's ability to identify cancer from x-ray images with 84% accuracy (Khan et al., 2018). Various ResNet architectures have demonstrated the ability to semantically segment satellite imagery with accuracy over 80%, depending on the specific architecture (Heryadi et al., 2020). While it is accepted that deeper neural networks tend to perform better under some circumstances (He et al., 2016a; Lodhi and Kang, 2019; Sinha et al., 2020), there can be benefits to using a shallower networks in some scenarios (Sekiyama et al., 2018; Gorban, Mirkes, and Tyukin, 2020).

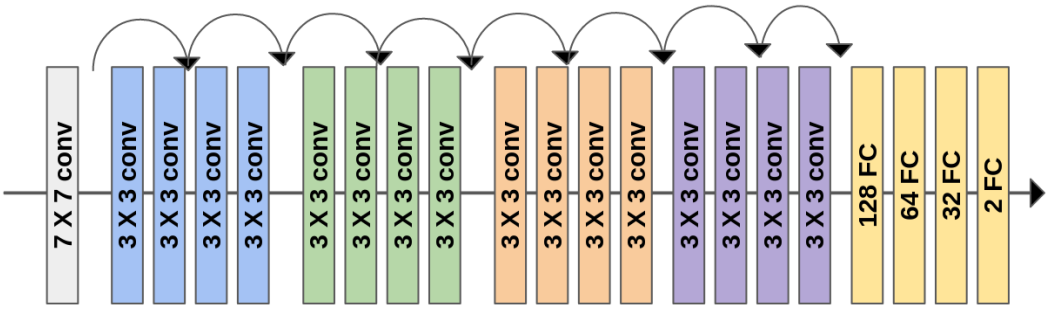


FIGURE 2.5: ResNet18, highlighting stacked convolutional layers, with residual learning depicted between intermediate convolutional layers.

2.4 Explainability

It is not uncommon for deep neural networks to be refereed to as "black boxes" (Dabkowski and Gal, 2017; Fong and Vedaldi, 2017; Petsiuk, Das, and Saenko, 2018; Chattopadhay et al., 2018; Naidu et al., 2020), due to the challenge of understanding why a model performs the way it does. This is a particularly notable challenge in the field of computer vision, where an algorithm may leverage unexpected features to distinguish between cases (Wei Tan et al., 2018).

There have been multiple attempts to provide semantically-interpretable descriptions of what features are most important within a target image; these include analyzing intermediate convolutional layers (Zeiler and Fergus, 2014), inverting convolutional networks to understand image salience (Mahendran and Vedaldi, 2015; Dosovitskiy and Brox, 2016), and Class Activation Mapping (CAM) (Zhou et al., 2016). Grad-CAM, introduced in (Selvaraju et al., 2017), was able to generalize CAM for use in more deep learning models, by analyzing the gradients in the last convolutional layer to create a visualization indicating what regions in the original image are important for classification. An example of Grad-CAM is displayed in figure 2.6.

Grad-CAM has lead to numerous variations with various benefits for specific scenarios. Grad-CAM++ is an attempt to provide better localization of objects in the scene for classification (Chattopadhay et al., 2018). Score-weighted CAM (Score-CAM) is a variant that avoids using gradients, and instead uses weights to determine importance in the image (Wang et al., 2020). Score-CAMpp uses

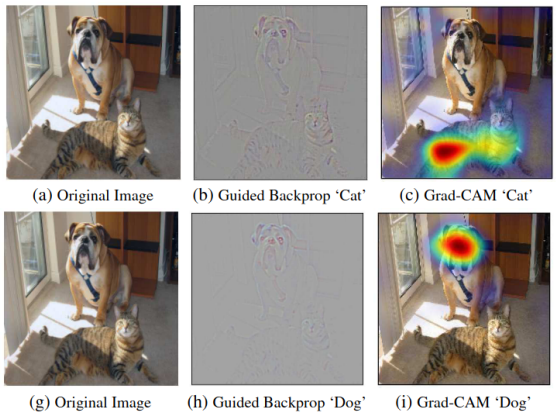


FIGURE 2.6: Images displaying the results of Grad-CAM analysis. This example shows the ability of CAM techniques to highlight the relevant portion of images used in classification. Image taken from (Selvaraju et al., 2017)

logarithmic transformation in an attempt to reduce non-target information, providing a more focused map (Shi et al., 2022). SSGrad-CAM is a spatial sensitive version of Grad-CAM that improves on object localization (Yamauchi and Ishikawa, 2022). Finally, Integrated Grad-CAM uses path integrals to better understand the regions of importance in a class activation mapping (Sattarzadeh et al., 2021).

Today, scholarship that explores the topic of explainability in the context of satellite imagery is sparse. Making the topic at hand more challenging, these past pieces have largely focused on discrete-case identification - for example, classifying well-defined semantic objects such as "airports", and identifying if the model identifies features such as "airplanes" (Vasu, Rahman, and Savakis, 2018; Charuchinda et al., 2019; Fu et al., 2019). This goal is distinct from the one posed by this dissertation, in which we seek to understand multiple urban features that may be correlated with an ill-defined semantic object ('protests' or 'riots'). This is challenging because of the following gaps in the literature: **1.** There are multi-target limits in exlainability techniques developed in conjunction with object based images. **2.** Satellite imagery is inherently less variable in band-space than other types of images. **3.** It is not immediately clear what is and is not semantically useful in satellite imagery-based classification.

The first key gap is multi-target limits on existing methods. These are cases in which multiple, discrete features that are not contiguous may be important

to classification to different degrees. For example, object detection in satellite imagery can struggle when there are many objects present, such as multiple airplanes at an airport (Tahir et al., 2022). Existing models struggle with this because often times the scale of objects in satellite imagery are smaller than objects in traditional images (Tahir et al., 2022) or if the model uses gradients to determine importance, the gradient can be relatively noisy (Wang et al., 2020).

The second challenge is in variability in band-space. Unlike traditional images, satellite imagery often has limited variability in pixel values (Carleer, Debeir, and Wolff, 2005). Due to the correlative nature of the spatial information present in geographic data, this variability can be overcome with higher resolution. However, with the introduction of increased resolution, segmentation can become more challenging (Carleer, Debeir, and Wolff, 2005). Furthermore, the band range in visible colors is smaller in satellite imagery, when compared to traditional images (Brewer, Lin, and Runfola, 2022). Having fewer differences in the imagery makes explaining the important features in the image more difficult.

The third difference is semantic utility. In traditional image recognition space, the semantic definitions of object are often well understood - i.e., the ear of a cat is easily interpret-able. In our context, a group of pixels that overlap a building, road and nearby park may be highlighted, leading to potential confusion over which feature (the building, road, or park) or combination of features was important to classification. An example of this might be only having a few pixels to available to detect and label an airplane (Tahir et al., 2022), as opposed to having many pixels extract features that can assist in classifying fish species (Rodrigues et al., 2010).

To overcome these limitations, in this dissertation, I propose to build on the capabilities of a technique called Score-CAM (Wang et al., 2020). Score-CAM is a CAM method that attempts to explain, with a human interpret-able visual display, the features within an image that determine classification. Score-CAM differs from traditional CAM methods that utilize gradients, and instead uses the forward pass scores of activation maps to determine the significance for target classes(Wang et al., 2020). When masking the highlighted portions of CAM results, Score-CAM only decreased 31.5% in predicted probability of the target class, compared to 47.8% and 45.5% decreases in GradCAM and GradCAM++

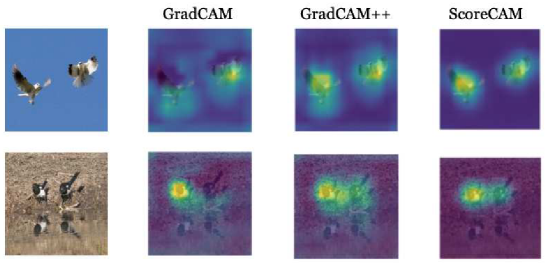


FIGURE 2.7: Images displaying the results of Score-CAM analysis. This examples highlight the performance of Score-CAM when there are multiple objects from a given class present in the image. Image taken from (Wang et al., 2020)

respectively (Wang et al., 2020). This average drop percentage illustrates Score-CAM's improved ability to correctly identify the portions of the image that are most pertinent to classification. Similarly, if the areas outside the CAM results are masked, Score-CAM increased in classification confidence by 30.6%, while GradCAM and GradCAM++ only increased by 19.6% and 18.9% respectively (Wang et al., 2020). This average increase percentage demonstrates Score-CAMs ability to focus on relevant portions of the image, not regions that do not assist in classification. For the purposes of this work, Wang et. al. found that it outperforms other techniques when there are multiple objects in a scene (Wang et al., 2020). Score-CAM was selected as a baseline model to develop further due to its ability to handle multiple objects, and our data are satellite images which are scene-centric, containing multiple objects. A figure displaying an example from the paper that introduced Score-CAM is shown in figure 2.7.

Chapter 3

Quantitative Analysis

3.1 Data & Methods

The primary objective of this work is to predict if a riot or protest will occur in a specific urban area, based solely on data from satellite imagery. In order to accomplish this objective, we leverage convolutional neural networks in combination with two data sources, ACLED (ACLED, 2022) and Planet (Planet, 2022a). We use these data to generate two different sets of information: the first set is satellite imagery of locations where riots occurred, and the second is a set of images of proximate areas (within the same city) that did not experience a riot event. Our deep learning model then seeks to disambiguate between these two cases, based on satellite imagery alone. This section provides details of our data processing and analytic approach.

3.1.1 Data

3.1.1.1 Selecting Riot Locations

Determining the locations where riots and protests have occurred is the first step in developing a data set for this work. To identify these locations, we leverage the Armed Conflict Location Event Data Project (ACLED), an open source database which contains information on a wide range of conflict types from across the globe (ACLED, 2022). ACLED contains more than 1.5 million events from 1997 to 2023, which are aggregated, categorized, and curated to create a data source that can specify time and location for conflict. We filter this database according to a number of criteria:

1. **Type of event.** We focus our analysis on protests and riots, which primarily represent urban unrest.
2. **Date.** We only leverage protest or riot events with a known date of occurrence.
3. **Geography.** Only events with a known neighborhood-level geographic footprint are selected.[1]

After filtering events, we are left with a resultant database of 53,307 events. In order to prevent over representation of any single unique location in the database, a maximum of 500 events were randomly selected from each neighborhood (i.e., "Seoul - Jongno"). After this stage, a total of 37,728 events across 1,089 unique locations were leveraged to construct our dataset of the location of conflict events.

3.1.1.2 Satellite Data

Once we have identified the location of riot events, we retrieve relevant *Planetscope* satellite imagery both (a) 24-48 hours prior to each event, and (b) in similar, nearby geographic locations that did not experience unrest. Planetscope - an integrated collection of images from the Dove, Dove-R and SuperDove satellites - provides four-band (RGB and NIR), approximately three to four meter spatial resolution satellite imagery with a daily temporal resolution (Planet, 2022b; see table 3.1). For both cases of imagery (with and without riot), we consider images that contain less than 50% cloud cover. An example of the imagery available can be seen in figure 3.1.

For each of the 37,728 instances of riots in our filtered ACLED dataset, we first retrieve a full scene of imagery from 24-48 hours prior to the event. These scenes are guaranteed to encompass the latitude and longitude representing the centroid of the neighborhood at which a conflict occurred; in cases where multiple images were available for a given event we chose the one closest in time to

[1]For example, some riots are known to have occurred in Beriut, while others occurred within neighborhoods in Beriut. There are 12 neighborhoods listed within some of the ALCED entries for Beriut (Ras Beirut, Port, Mazraa, Achrafieh, Mousseitbeh, Saifi, Minet El Hosn, Rmeil, Bachoura, Medawar, Ain Mreisseh, and Zokak El Blat). These neighborhood specific entries have neighborhood specific latitudes and longitudes, and we use these neighborhood specific events to construct our data set.

Band	**Dove Classic** Wavelength (nm)	**Dove-R** Wavelength (nm)	**SuperDove** Wavelength (nm)
Red	590 - 670	650 - 682	650 - 680
Green	500 - 590	547 - 585	547 - 583
Blue	455 - 515	464 - 517	465 - 515

TABLE 3.1: Technical wavelength specifications for RGB bands of Planetscope sensors (Planet, 2022b).

FIGURE 3.1: Satellite Image of Athens Greece, taken 31 January 2018. Imagery ©Planet Labs PBC 2023. All rights reserved.

TABLE 3.2: PlanetScope Constellation (Planet, 2022b)

Instrument	**Image area**	**Availability**
Dove Classic	25 x 11.5 sq km	July 2014 - April 2022
Dove-R	25 x 23 sq km	March 2019 - April 2022
SuperDove	32.5 x 19.6 sq km	March 2020 - present

the event (with a minimum of 24 hours prior to the event). Ultimately, this process resulted in 19,902 satellite scenes being downloaded, with an average spatial dimension that can vary depending on the generation of satellite (see table 3.2) and geographic latitude of collection. Because riots may occur at the same location, but at multiple points in time, some locations (i.e., a seat of government, culturally significant locations, etc) may appear in the database multiple times; the most common of these occurrences are summarized in table 3.3.

Country	**Neighborhood**	**Count**	**Earliest Date**	**Latest Date**
Pakistan	Karachi - Saddar	278	7 October 2017	30 September 2022
Iran	Tehran - District 6	270	9 October 2017	26 September 2022
Iran	Tehran - District 12	268	9 October 2017	28 September 2022
Lebanon	Beirut - Port	252	7 October 2017	26 September 2022
Greece	Athens - Central Athens	247	18 January 2018	28 September 2022
South Korea	Seoul - Jongno	240	18 January 2018	21 September 2022
South Korea	Seoul - Jung	226	18 January 2018	26 September 2022
Italy	Rome - City Center	222	7 January 2018	23 September 2022
India	Delhi - New Delhi	220	2 October 2017	4 September 2022
South Korea	Seoul - Seocho	220	8 January 2018	28 September 2022

TABLE 3.3: Neighborhood locations which occur most frequently. 'Earliest' and 'Latest' date refer to the earliest and latest date of a protest event for each neighborhood. For example, in the neighborhood of Seocho in Seoul, 220 independent protest or riot events occurred from January 8th, 2018 to September 28th, 2022. In our analysis, this would be represented by 220 individual satellite tiles, each taken between 24 and 48 hours before the actual event.

From the satellite scene retrieved for each conflict event, we extract two types of data. First, we extract a one kilometer by one kilometer box centered on the conflict event neighborhood. This box is saved and identified as the location of the unrest in our database.

Second, we extract a number of cases to serve as null events - i.e., locations from the same urban area, but where no unrest occurred. To generate these null cases, we follow a multiple step process in which we:

1. **Identify urban areas.** We only consider areas in the scene that have a population density over 300 inhabitants per kilometer.

2. **Exclude areas that are within 10km of the conflict event.** We isolate the conflict event by removing the urban areas that are within ten kilometers of the centroid of the neighborhood in which conflict occurred.
3. **Sample.** With the remaining urban areas in the satellite scene, we generate a list of random centroids which are constrained to be a minimum of 2 kilometers apart, and select a maximum of 10 of these to generate 1km box 'null' locations at which no protest or conflict occurred. The 2 kilometer separation ensures that none of our null boxes overlap.

In step one, we overlay information about the degree of urbanization (Schiavina, Melchiorri, and Pesaresi, 2023; European Commission and Statistical Office of the European Union, 2021) onto each satellite scene to determine what portions are urban, and which parts are not. This is accomplished by using the DEGURB dataset (Schiavina, Melchiorri, and Pesaresi, 2023), which was developed by the European Commission's Joint Research Centre. This data categorizes geographical areas into Urban Centre, Urban Clusters (including towns and suburbs), and Rural Grid Cells (including villages and dispersed rural) zones based on population density and contiguity of dense areas (European Commission and Statistical Office of the European Union, 2021). Its creation involves analyzing high-resolution satellite imagery and detailed population survey data, with the goal of providing an accurate representation of the urban landscape with one kilometer spatial resolution (European Commission and Statistical Office of the European Union, 2021). The DEGURB dataset used in this work is representative of 2020 (Schiavina, Melchiorri, and Pesaresi, 2023). For our purposes we consider anything with a density over 300 inhabitants per square kilometer as urban (European Commission and Statistical Office of the European Union, 2021). The results of this approach are illustrated in figure 3.2. This binary representation of urban areas is then be applied to each satellite scene as a mask, allowing us to select null cases from proximate urban areas.

In step two, in order to ensure the areas selected for null cases are distinct from the areas of unrest, we exclude all urban areas up to 10 kilometers away from the centroid of the riot neighborhood from consideration, as illustrated in figure 3.3.

Third, after excluding the ten kilometer region around each unrest event,

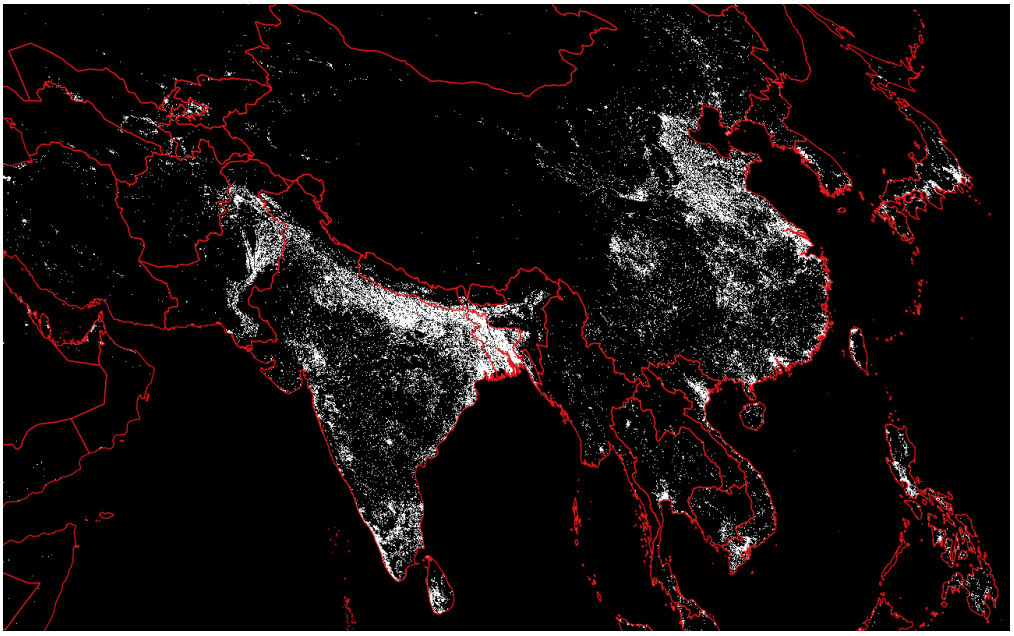


FIGURE 3.2: A portion of the DEGURB data, highlighting areas of the world that are considered urban in our data set. DEGURB defines urban regions as those with a density more than 300 inhabitants per km European Commission and Statistical Office of the European Union, 2021. Red lines represent country level boundaries (Runfola et al., 2020).

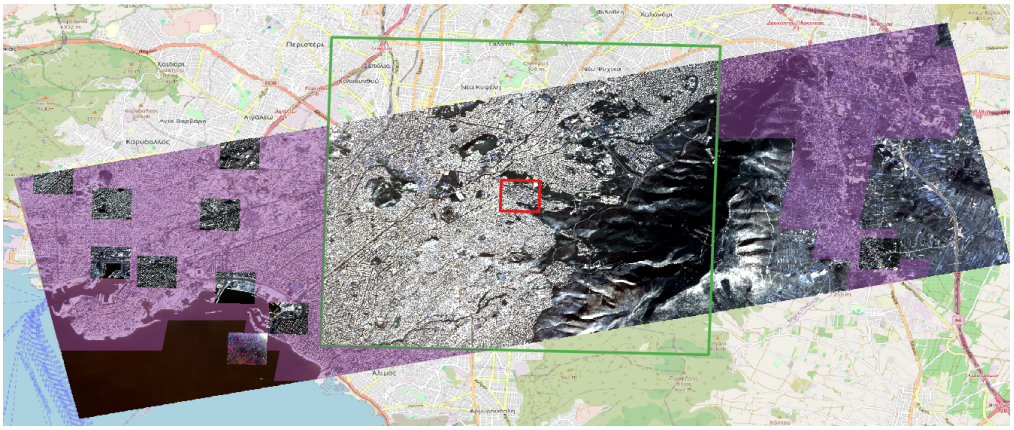


FIGURE 3.3: Satellite Image of Athens Greece, taken 31 January 2018. The red box in the center of the image is a 1 kilometer box around the riot location. The green box is a 10 kilometer exclusionary area around the riot location, from which we do not draw "null" case contrasts. Areas which fall outside the green box, that are also urban, are eligible for selection (displayed in purple). From the potential null region, we sample random, non-overlapping 1 kilometer boxes to generate null location clips. Imagery ©Planet Labs PBC 2023. All rights reserved.

from the remaining urban regions in the satellite scene we select random locations for null-riots. We accomplish this by generating a list of random latitudes and longitudes that are within the available regions. We ensure that each of these random locations are at least 2 kilometers away from any other locations on our random list. We then take a maximum of ten of these locations, and construct a 1 kilometer box around each one. We construct up to 10 null cases (that do not overlap) from the eligible urban regions from each scene (noting that less dense urban areas are occasionally represented by less than 10 null cases due to a lack of proximate urban areas). A visualization of the results from this process can be seen in figure 3.3.

After this process is completed, for each conflict event we are left with a set of one ($1km^2$) kilometer box representative of where unrest occurred, and up to 10 ($1km^2$) km boxes representative of urban areas proximate to the unrest event, but with no known activity. Across our full dataset of 19,902 unrest locations, 18,634 (93.6%) had 10 null cases available; the distribution of null cases across images can be seen in figure 3.10. Our final dataset included only locations that had the full complement of null clips, for a total of 18,631 cases of unrest, and

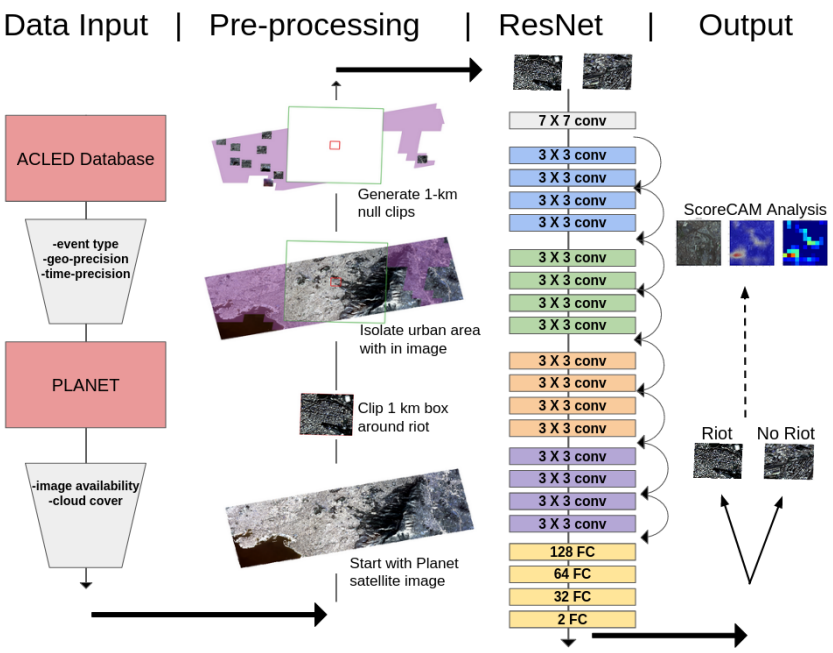


FIGURE 3.4: A synopsis of our overall modeling architecture. Stages include the collection of data, pre-processing, network training, categorization, and explainability analysis. Imagery ©Planet Labs PBC 2023. All rights reserved.

186,310 null cases. We then normalize all of these image clips based on a sample of the full satellite scenes (Goodman, BenYishay, and Runfola, 2021; Runfola et al., 2022; Lv et al., 2024; Brewer, Lv, and Runfola, 2023). Tests of different permutations of this dataset (i.e., models with a 1:1 ratio of null and riot cases) can be found in section 3.4.1.

3.1.2 Methods

Our overall modeling architecture is summarized in figure 3.4. In order to estimate the likelihood of if an unrest event occurred or not at each location, we leverage a ResNet18 (He et al., 2016a) as our base model, replacing the fully connected layer with a series of dense layers with 128, 64, and 32 hidden nodes. In order to improve the efficiency of our training, following other literature in the satellite imagery analysis space (Goodman, BenYishay, and Runfola, 2021; Brewer et al., 2021; Runfola et al., 2022; Lv et al., 2024; Brewer, Lv, and Runfola, 2023), pre-trained weights from ImageNet were used as our initial baseline.

Model	Learning Rate	L2 Decay	Freeze Layers	Drop Out	Test Acc	TP	FP	FN	TN	Precision	Recall	F1
A	0.000001	None	None	No	92.5%	56	34	122	1,861	62.2%	31.5%	41.8%
B	0.000015	0.01	First 5	No	90.5%	95	103	93	1,782	48.0%	50.5%	49.2%
C	0.00001	0.001	First 5	No	92.2%	95	55	107	1,816	63.3%	47.0%	54.0%

TABLE 3.4: Representative results from hyperparameter tuning efforts. All training iterations were based on the same ResNet18 architecture, training with the same 1,000 satellite images from the full dataset, for 40 epochs.

3.1.2.1 Hyperparameter Search

Prior to training on all 18,631 images, we first randomly selected a subset of 1,000 conflict events (1,000 unrest cases and 10,000 null cases) to implement a grid search across hyper-parameters [2]. To account for class imbalance, we implement a weighted cross entropy loss (Ho and Wookey, 2019) with an ADAM optimizer (Kingma and Ba, 2014) for our training procedure.

Our hyperparameter search encompassed trials of different learning rates, L2 regularization, dropout, freezing layers (results and parameters from a sample of the trials can be seen in the appendix in section 3.4.2). Results from a selection of three of the best performing cases in the hyperparameter testing are shown in table 3.4. On the basis of these results, we selected one model (denoted as Model C in table 3.4) to test on the full dataset, which is described in table 3.5.

We assess our model by interpreting the overall accuracy, precision and recall. The precision is the ratio of true positives to the number of positive predictions our model made (Davis and Goadrich, 2006), which will measure our model's ability to correctly predict riots when it does makes a prediction. The recall is the ratio of true positives to the number of riots in the data set, which measures our model's ability to identify how frequently riots are occurring (Davis and Goadrich, 2006).

3.1.2.2 Additional Analyses

In addition to identifying the best convolutional model performance, we additionally implement two additional analyses to better understand the strengths

[2]Training was performed using pyTorch on 8 RTX 6000 NVIDIA GPUs. On average, models trained using the hyperparameter dataset took approximately 6.5 hours to complete 40 epochs; our full model across all images took 321 hours for 100 epochs.

and weaknesses of this approach. These include (a) generating information on the country-level performance of the model, and (b) and explanatory model that sought to identify the features within a given image that were correlated with conflict events (or the lack thereof).

To explore the spatial distribution of accuracy of the approach, we first filter our data to only consider countries that had 500 or more observations (a minimum of 250 riot clips and 250 null clips). This created a validation set consisting of 32,548 clipped images, distributed across 24 countries (see table 3.6). From this, we withhold 20% of each country's observations for validation after training. This ensures that each country has at least 100 observations (50 riot clips and 50 null clips) for validation. We then selected the hyperparameters from our best performing model (model C, see table 3.4), and trained a ResNet18 using 80% of the validation data (26,058 images, half riot or protest and half null) for 50 epochs. We then used the withheld 20% of images (6,490 images, half riot or protest and half null) to test for accuracy within each country.

To begin to explore the underlying drivers of model performance, we additionally take preliminary steps towards trying to assess what features the model may be identifying and using in predictions. To implement this process, we leverage Score-CAM (Wang et al., 2020). Score-CAM is a Class Activation Mapping (CAM) method that attempts to explain, with a human interpret-able visual display, the features within an image that determine classification. Score-CAM differs from traditional CAM methods that utilize gradients, and instead uses the forward pass scores of activation maps to determine the significance for target classes (Wang et al., 2020). For the purposes of this work, Wang et. al. found that it outperforms other techniques when there are multiple objects of relevance in the scene (Wang et al., 2020), a nearly universal characteristic of satellite imagery.

Test Accuracy	**97.39%**
True Positives (predict riot)	2,741
False Positives	163
False Negatives (missed riot)	905
True Negatives	37,180
Precision	94.39%
Recall	75.18%
F1 Score	83.69%

TABLE 3.5: Results from ResNet18 using the full data set.

3.2 Results

3.2.1 Full Data Set

In this section, we report our findings from our analysis of the full dataset (N= 204,941 clipped images), using the best performing model from our hyper-parameter testing (model C, as described in table 3.4). Results of this model are presented in table 3.5.

As table 3.5 shows, the approach outlined in this paper achieve an overall accuracy of 97.39% - i.e., of the 40,989 images in the test dataset, 39,921 were correctly identified in terms of if the image is likely to be the site of a protest event or not.[3] There are 3,646 riot or protest images in the testing set and the model correctly identifies 2,741 of these, resulting in a recall score of 75.18%. This demonstrates the model's ability to distinguish riot/protest events from non-riot events. The model predicts there will be a riot in 2,904 of the images and is only incorrect 163 times producing a precision score of 94.30%. In the context of our scenario, when the model predicts there will be a riot or protest in an image, it is correct over 94% of the time.

In addition to the global accuracy, we also subset our data by country and report accuracy within each based on a validation set (see section 3.1.2.2 of our methods). The results of this country-specific validation testing are shown in

[3]It is important to note that our data set is constructed in a manner that would result in relatively high test accuracy. We have one riot and ten null riot clips per satellite scene. This means that if our model predicted no riot for every clipped image, the model would be correct 90.9% of the time. Even given imbalance in the data set, our trained model achieves better results, accurately predicting riots and null riots over 97% of the time. Further explorations of the value of the model in the context of imbalance are described in section 3.4.1 of the appendix.

Country	**Images**	**Country**	**Images**
South Korea	7,494	South Africa	1,480
Pakistan	2,622	Chile	1,302
Iran	2,334	Japan	1,256
Lebanon	1,656	India	1,148
Palestine	1,572	Brazil	1,112
China	1,550	Bangladesh	1,092
Country	**Images**	**Country**	**Images**
Ukraine	924	Greece	634
Thailand	890	Yemen	604
Italy	728	United Kingdom	566
Indonesia	678	Taiwan	562
Russia	668	Peru	522
Venezuela	648	Iraq	506

TABLE 3.6: There are 32,548 clipped images in the validation data set. Half of these are from riots/protests, and half are null clips. Only countries that have at least 500 images are included. 20% of each county's images will be withheld from training and testing, and used in validation.

table 3.8.[4] Lebanon (94.5%), Iran (94.4%), and Pakistan (92.6%) were the most accurate in this analysis, while Yemen (78.3%), Russia (78.0%), and Peru (77.9%) were the least accurate countries. No clear regional patterns existed, though some evidence suggests that accuracy and total number of observations may be correlated (i.e., less accurate news media reporting in Russia may be attributable to the lower accuracy in that context).

Of note, we observe a strong correlation between our softmax classification scores and accuracy within each country around the world, suggesting that softmax scores can be used as a proxy for prediction confidence (see figure 3.6). While softmax may bias towards higher degrees of confidence (Pearce, Brintrup, and Zhu, 2021; Subramanya, Srinivas, and Babu, 2017), as a relative metric

[4]Of note, the re-trained model which withheld data for each individual country had a slightly lower global accuracy than our full results, of 89%. While this 89% accuracy is lower than the accuracy from the full data set shown in table 3.5, the testing circumstances of the validation are more challenging due to the even split between riot and null in the validation data set. Additionally, there are known limitations in our images that likely contribute to a decrease in performace. These limitations are discussed in greater depth in section 4.1.3.1. The precision and recall for the validation set were inline with the accuracy, each scoring slightly over 89%.

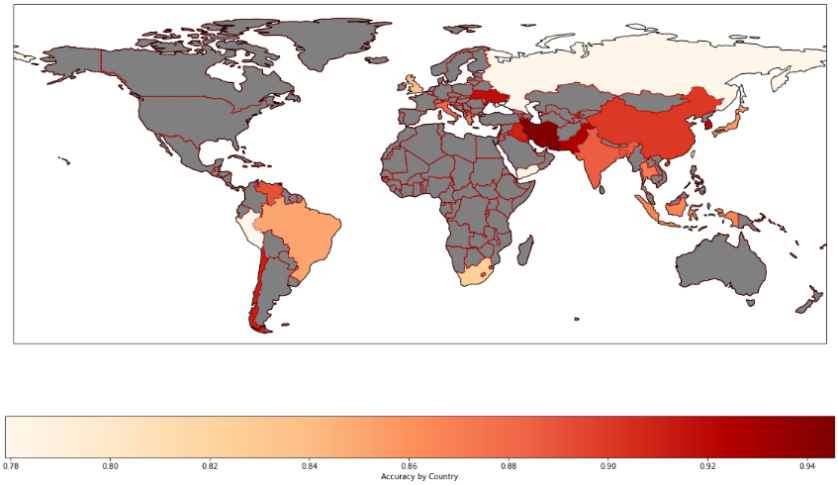


FIGURE 3.5: Map of 24 countries included in validation testing. Each country in the validation testing has a minimum of 100 images.

Test Accuracy	**89.41%**
True Positives (predict riot)	2,903
False Positives	345
False Negatives (missed riot)	342
True Negatives	2,900
Precision	89.38%
Recall	89.46%
F1 Score	89.42%

TABLE 3.7: Results from validation testing.

Country	**Count**	**Accuracy**	**TP**	**FP**	**TN**	**FN**
Lebanon	330	94.54%	159	12	153	6
Iran	466	94.42%	225	18	215	8
Pakistan	524	92.56%	247	24	238	15
South Korea	1498	92.12%	700	69	680	49
Ukraine	184	91.85%	84	7	85	8
Chile	260	91.15%	120	13	117	10
Iraq	100	91.00%	45	4	46	5
China	310	90.00%	140	16	139	15
Palestine	314	89.49%	141	17	140	16
Venezuela	128	89.06%	58	8	56	6
Bangladesh	218	88.99%	90	5	104	19
India	228	88.60%	97	9	105	17
Italy	144	88.19%	62	7	65	10
Greece	126	87.30%	62	15	48	1
Thailand	178	87.08%	75	9	80	14
Indonesia	134	86.57%	62	13	54	5
Japan	250	85.60%	105	16	109	20
Brazil	222	85.14%	94	16	95	17
United Kingdom	112	83.93%	50	12	44	6
South Africa	296	82.43%	114	18	130	34
Taiwan	112	82.14%	45	9	47	11
Yemen	120	78.33%	41	7	53	19
Russia	132	78.03%	50	13	53	16
Peru	104	77.88%	37	8	44	15

TABLE 3.8: Results from country level accuracy after validation testing. These results are listed from highest accuracy to lowest accuracy. We have also included the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) for each country.

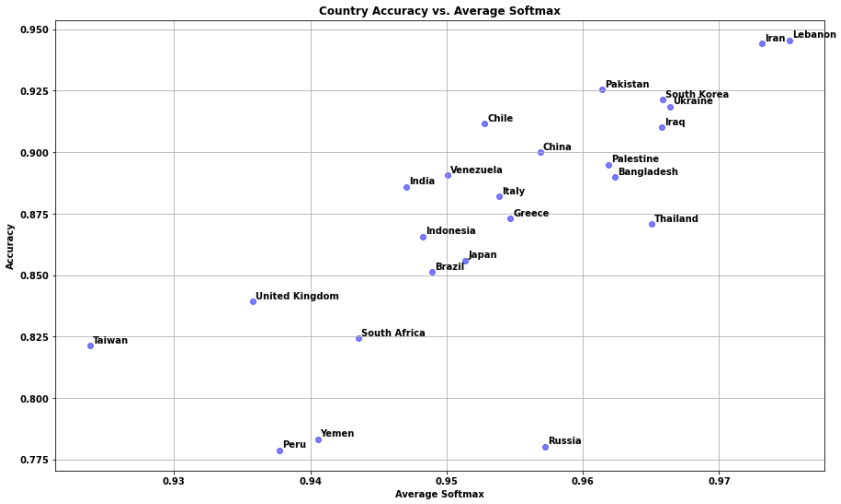


FIGURE 3.6: The average softmax for each country when compared to the average accuracy of prediction of each country. Of note, the axis's do not begin at 0, but instead focus in on the domain and range of the values in the data.

it may provide helpful guidance to policymakers seeking to use these types of methods.

3.2.2 Explainability of Results

For our best performing model (model C in the table 3.5), we implemented Score-CAM on a subset of randomly selected, paired locations, ultimately consisting of 1,089 riot locations, and 1,089 null locations. The Score-CAM results were then visually reviewed in an attempt to discern patterns in what the trained ResNet used in classification. Interpreting the results of Score-CAM is inherently qualitative, making this a rich area for future work; we discuss this limitation further in section 4.1.3.2.

Despite the limitations of Score-CAM, the visual interpretation indicated a few clear patterns. An example of the first of these is displayed in figure 3.7, in which we can observe a large sports stadium in the image in the southeast region. This large stadium is the location which Score-CAM identifies as the portion of the image which leads towards the classification (indicated through brighter values in the displayed heatmap). In this case, the sports stadium lead the ResNet to classify the scene as a location that is unlikely to experience a riot.

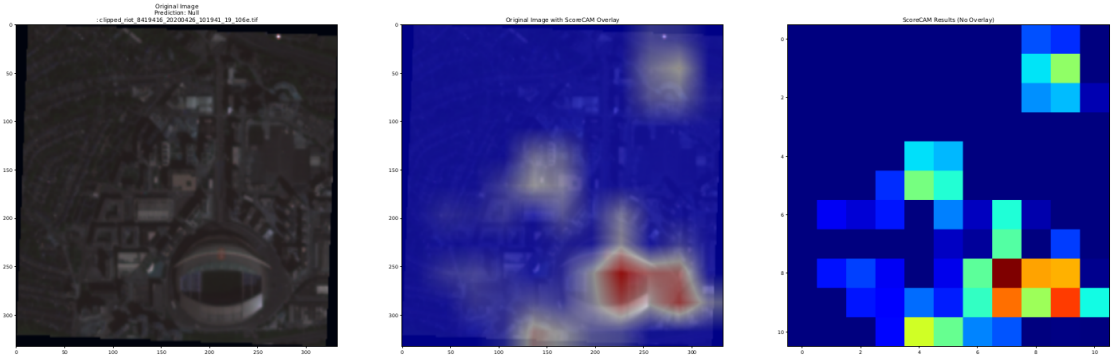


FIGURE 3.7: Example clipped image on the left. The clipped image, a one kilometer box around a riot location. The Score-CAM overlayed on top of the image is shown in the middle. The Score-CAM visual is displayed on the right. Imagery ©Planet Labs PBC 2023. All rights reserved.

We can see another example in figure 3.8, in which again, the ResNet identifies the sports stadium as the reason to classify the scene as a non-riot. We do not offer any explanation for why the sports stadiums are indicative of a non-riot scene, but these stadiums provide an example of the specific features which ResNet is learning to make classification decisions.

Another example highlighted in the Score-CAM analysis is shown in figure 3.9. We can see a densely populated area, with a large open park or green space in the center of the image. The trained network correctly predicted this image was from a riot or protest. When we reference the ACLED data, this image is from a protest in the Lalbagh neighborhood of Dhaka, Bangladesh. Lalbah is a fort built during the Mughal period in 1678, which was used subsequently by the British and Bangladesh governments as a location of governance and influence (Shakur, Islam, and Masood, 2010). Today, it is a location containing monument's and statues symbolizing rulers and regimes of the past, that is known as a common location for protests in the city of Dhaka (Begum, 2018). While the deep learning model was not aware of these historic contexts, the unusual land use and associated image features were sufficient to classify this as a likely location of riots.

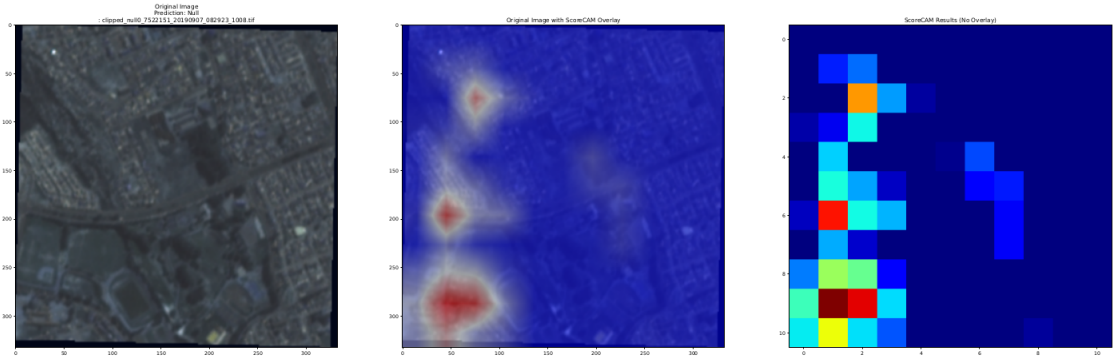


FIGURE 3.8: Example clipped image on the left. The clipped image, a one kilometer box around a non riot location. The Score-CAM overlayed on top of the image is shown in the middle. The Score-CAM visual is displayed on the right. Imagery ©Planet Labs PBC 2023. All rights reserved.

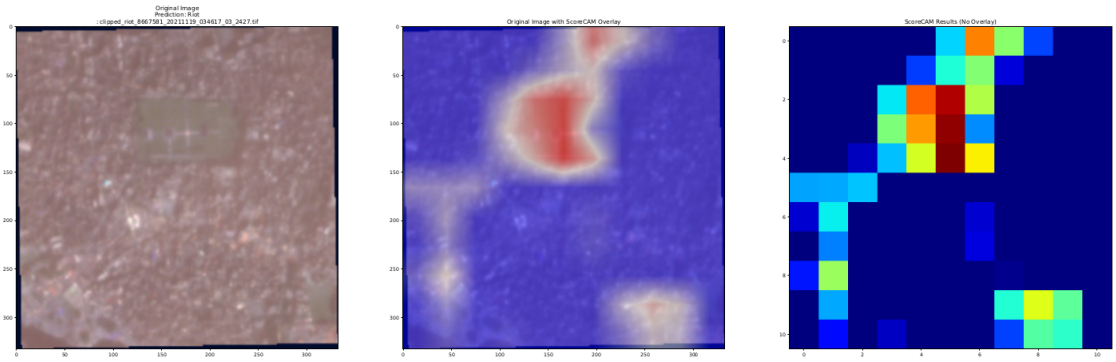


FIGURE 3.9: The image on the left is a centered on Lalbagh Fort in Dhaka Bangladesh, taken on 19 November 2021, lesss than 48 hours before a protest at that location. The Score-Cam visual is displayed on the right. Imagery ©Planet Labs PBC 2023. All rights reserved.

3.3 Discussion and Conclusions

The results presented in table 3.5 provide evidence that satellite information alone can provide useful information for the purposes of predicting where, within an urban environment, protests and riots are most likely to occur. While this finding is likely to be of interest to those operating in data-sparse environments, it is well supported by past social science literature highlighting the interconnected nature of urban form and social processes (Fox and Bell, 2016; Begum, 2018). By engaging in a global-scope study, here we are able to exploit this correlation by learning what these patterns are, and then leveraging them in estimation. This finding held true across multiple model and data permutations (see tables 3.5, 3.7 and 3.8), indicating that - even in some of the most challenging situations (i.e., relatively small training and validation sets), model accuracy can approach or exceed 90%.

Furthermore, this technique performs well across the globe. As highlighted in figure 3.5, there do not seem to be any regions that under perform. Many countries with a relative low accuracy score (i.e., Russia) are in close proximity to a country with a higher accuracy score (i.e., China). This pattern holds across the globe in South America, Asia, the Middle East, and Europe.

Of note, in our softmax analysis seeking to correlate scores to accuracy, a single outlier, Russia, is observed in figure 3.6 and table 3.8. Russia has a lower comparative accuracy to other countries with similar softmax results. This might be indicative of Russia's control of news sources (Gehlbach, 2010), or inherit bias in ACLED's collection of data which relies on news sources and non-governmental observation organizations that might not be focused on Russia.

3.3.1 Conclusions

In this work we constructed a data set consisting of 204,941 satellite images of riots and protests across the world. After subseting the images into two classes of riots and non-riots, we trained a ResNet18 to identify which images were from locations associated with a riot. When fine-tuned, our model achieved an accuracy of over 97%, suggesting that satellite imagery has information of relevance and value to estimating the location of riot events. This was true across a wide range of different tests and permutations of the data. We further provide some

initial exploration into the explainability of this model, leveraging ScoreCAM to identify features the model is using in the classification task.

3.4 Supplemental Information

3.4.1 Deduplication Tests

In this section, we present a test that controls for both class imbalance and geographic bias in our data. Our methodology leverages a large set of training data, specifically relying on an arbitrary 10:1 ratio of 10 null cases (no conflict event) to 1 positive case (a location where a conflict occurred). Furthermore, some geographic locations are in the database multiple times - i.e., there may have been multiple protests at the same geographic location, even if they are at different dates (see table 3.3). This results in both class imbalance (10 null cases for every 1 positive case), and geographic biases in where events are drawn from. The class imbalance will potentially inflate accuracy scores, given a 10 to 1 ratio of null clips to riot clips - i.e., an untrained model could simply predict null for all images, and achieve an accuracy of 90.9%. Additionally, with repeated locations, the model will see the riot clip locations multiple times (i.e., even when each satellite scene has unique spatial information as it is drawn from a different date, the 1-km box centered on the latitude and longitude of the neighborhood will be the same). This might allow our network to learn the specifics of a location, and over-fit to particular locations, instead of learning what features in urban areas predict riots and protests. Therefore we constructed a limited data set to control for these issues.

To test if these attributes of our data resulted in bias, a new dataset was constructed which limited the data to a single riot image (1,089 1-km boxes) and a single non-riot image (1,089 1-km boxes) per location. This means that our model was only able to analyze a riot location a single time during training, regardless of how frequently riots might happen at that location. This should be a much harder training task for the model, with far less data available (2,178 images in total; these 2,178 images represent roughly 1% of the data available for training in the full data set of 204,941 images). Under these constraints, the maximum classification accuracy we observed was 67.37% (see table 3.9). Of

Test Accuracy	**65.37%**
True Positives (predict riot)	154
False Positives	105
False Negatives (missed riot)	46
True Negatives	131
Precision	59.46%
Recall	77.0%
F1 Score	67.1%

TABLE 3.9: Results from ResNet18 using only a single riot clip and single null riot clip per location.

note, the recall scores for our full data set and limited data set were very similar (75.18% and 77.0% respectively), despite the different size and scope of the training data.

These results suggest that - even under extremely challenging, small-N circumstances - deep learning models can still identify meaningful features that are correlated with protest and riot events from satellite imagery.

3.4.2 All Results

While we focus on our best performing models throughout this piece, there were a number of additional permutations and tests we performed while identifying the best modeling strategies, which we present here. We began grid search across select hyperparameters, using a small test set of 100 random samples from our full data set. Initially we were concerned with narrowing down the selection of the best performing learning rates, freezing layers of the ResNet, and dropping out connections between our fully connected layers. The results of a sample of these are shown in table 3.10 and table 3.11.

After the initial grid search, we increased the size of data set to 1,000 locations (1,000 riot clips, and 10,000 null clips). We also adjusted the hyperparameter grid search space. Our best performing model referred to as Model C in table 3.4, is Config 10 in table 3.13. Config 10 has the highest F1 Score across these grid search results, reflecting the best balance between Precision and Recall. Due to this strong performance, these parameters were used to train the full dataset.

Metric	A1	A2	B1	B2	C1	C2
Test Accuracy (%)	91.59	91.12	90.65	93.93	93.46	86.45
True Positives	0	0	0	0	0	0
False Positives	0	0	4	0	0	0
False Negatives	18	19	16	13	14	29
True Negatives	196	195	194	201	200	185
Precision (%)	0.00	0.00	0.00	0.00	0.00	0.00
Recall (%)	0.00	0.00	0.00	0.00	0.00	0.00
F1 Score (%)	0.00	0.00	0.00	0.00	0.00	0.00
Learning Rate	1e-06	1e-06	1e-06	1e-06	1e-06	1e-06
Freeze Layers	0	0	5	5	10	10
Drop Out Pair	(0, 0)	(0.1, 0.05)	(0, 0)	(0.1, 0.05)	(0, 0)	(0.1, 0.05)
L2 Weight Decay	0	0	0	0	0	0

TABLE 3.10: All of the models in this table were tested with 100 random locations (100 riot clips and 1,000 null clips). In this table, all models used a learning rate of 1e-06. Models froze either none of the ResNet layers (A1, A2), the first 5 layers (B1, B2), or the first 10 layers (C1, C2). Between the first two and the second two layers, none of the connections were dropped (A1, B1, C1), or 10% and 5% were dropped (A2, B2, C2).

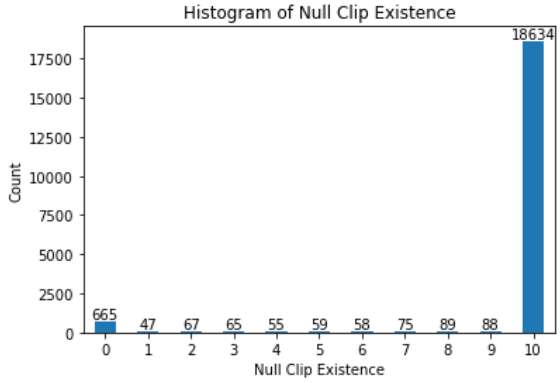


FIGURE 3.10: Distribution of null clips from the full 19,902 images downloaded. Instances where less than 10 clips were taken are primarily due to the amount of urban area available in the satellite image. There were three additional locations that were eventually able to provide 10 null clips, but not included before the dataset was finalized with 18,631 locations at training time.

Metric	D1	D2	E1	E2	F1	F2
Test Accuracy (%)	88.78	86.92	91.12	91.59	88.32	88.78
True Positives	5	1	7	5	0	0
False Positives	7	13	4	6	0	0
False Negatives	17	15	15	12	25	24
True Negatives	185	185	188	191	189	190
Precision (%)	41.67	7.14	63.64	45.45	0.00	0.00
Recall (%)	22.73	6.25	31.82	29.41	0.00	0.00
F1 Score (%)	29.41	6.67	42.42	35.71	0.00	0.00
Learning Rate	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05
Freeze Layers	0	0	5	5	10	10
Drop Out Pair	(0, 0)	(0.1, 0.05)	(0, 0)	(0.1, 0.05)	(0, 0)	(0.1, 0.05)
L2 Weight Decay	0	0	0	0	0	0

TABLE 3.11: All of the models in this table were tested with 100 random locations (100 riot clips and 1,000 null clips). In this table, all models used a learning rate of 1e-05. Models froze either none of the ResNet layers (D1, D2), the first 5 layers (E1, E2), or the first 10 layers (F1, F2). Between the first two and the second two layers, none of the connections were dropped (D1, E1, F1), or 10% and 5% were dropped (D2, E2, F2).

Metric	Config 1	Config 2	Config 3	Config 4	Config 5	Config 6
Test Accuracy (%)	91.80	91.80	91.27	90.69	92.52	91.85
True Positives	62	39	43	77	25	78
False Positives	54	11	41	65	7	62
False Negatives	116	159	140	128	148	107
True Negatives	1,841	1,864	1,849	1,803	1,893	1,826
Precision	0.5345	0.7800	0.5119	0.5423	0.7812	0.5571
Recall	0.3483	0.1970	0.2350	0.3756	0.1445	0.4216
F1 Score	0.4218	0.3145	0.3221	0.4438	0.2439	0.4800
Learning Rate	1e-05					
L2 Weight Decay	0.1			0.01		
Freeze Layer	0	0	0	5	5	5
Dropout Pair	(0, 0)	(0.1, 0.05)	(0.5, 0.1)	(0, 0)	(0.1, 0.05)	(0.5, 0.1)

TABLE 3.12: Model performance metrics for configurations 1 to 6 with learning rate of 1e-05, with variations in L2 weight decay, freeze layer, and dropout pair settings.

Metric	Config 7	Config 8	Config 9	Config 10	Config 11	Config 12
Test Accuracy (%)	91.51	89.77	92.91	92.33	90.16	92.76
True Positives	85	86	54	86	92	80
False Positives	64	106	34	32	101	49
False Negatives	112	106	113	127	103	101
True Negatives	1,812	1,775	1,872	1,828	1,777	1,843
Precision	0.5705	0.4479	0.6136	0.7288	0.4767	0.6202
Recall	0.4315	0.4479	0.3234	0.4038	0.4718	0.4420
F1 Score	0.4913	0.4479	0.4235	0.5196	0.4742	0.5161
Learning Rate	1e-05					
L2 Weight Decay	0.01			0.001		
Freeze Layer	0	0	0	5	5	5
Dropout Pair	(0, 0)	(0.1, 0.05)	(0.5, 0.1)	(0, 0)	(0.1, 0.05)	(0.5, 0.1)

TABLE 3.13: Model performance metrics for configurations 7 to 12 with learning rate of 1e-05, transitioning from L2 weight decay settings of 0.01 to 0.001, including variations in freeze layer and dropout pair settings.

Chapter 4

Chapter Roadmaps

4.1 Dissertation Paper 1: Conflict prediction using Satellite Imagery

In this chapter of my prospectus, I will outline each of my three dissertation papers, including my methods, data, and plans for implementation. My first paper, presented here, is predicated upon the results from my quantitative study, as shown in chapter 3. Because of overlap between my quantitative analysis and this paper, I provide a synopsis of each major stage of my work here, and refer you to chapter 3 for a broader discussion.

4.1.1 Major Research Question

Can satellite imagery alone be used to determine the likelihood of conflict in urban areas?

4.1.2 Proposed Data & Methods

4.1.2.1 Data

Our full data processing pipeline is presented in section 3.1.1, and briefly summarized here. In order to test this hypothesis, we first need to determine the locations where riots and protests occurred in the past. To accomplish this, we filter events from the full ACLED database using the following criteria: **1)** include riots and protests, **2)** an event with a known specific date, **3)** and events with a neighborhood-specific geographic footprint. This criteria generates a list of 53,307 riots and protests. We avoid over representing any single location, by

only allowing locations to appear a maximum of 500 times. This further filters our list of potential locations to 37,728 events.

With this filtered list, we attempt to download satellite images of these 37,728 riots or protests. Our strategy to download images is to search for satellite images from 24-48 hours prior to the riot or protest. Also, we only consider images that contain 50% or less cloud cover. Under these constraints, we download 19,902 satellite images. We will use these satellite images to construct our data set.

From each satellite image, we clip a 1 km^2 box centered on the latitude and longitude of the riot or protest. We label these clipped images as "riot" in our data set. The next task is to generate up to 10 null clips, each of which is a 1 km^2 box. For each satellite image we use the following criteria: **1)** null clips must be a minimum of 10 km away from the riot, **2)** null clips must be from urban regions of the satellite image, **3)** null clips are spaced far enough apart, so that none of the null clip 1 km^2 boxes overlap. We label these clipped images as "nulls" in our data set.

This generates a data set that consists of 1 km^2 boxes, that are labeled either "riot" or "null". Specifically, the data set contains 18,631 "riot" images, and 186,310 "null" images. In total, the data set has 204,941 images that we can use in training.

4.1.2.2 Methods

In order to determine the likelihood of a protest or riot, we train a ResNet18 convolutional neural network. Starting with a pre-trained ResNet18, we can train the network to predict if clipped images are from riots/protests or from null locations. Additionally, we explore the explainability of these clipped images using Score-CAM. We also subset our data to explore any region or cultural patterns that might emerge in our analysis. Our full implementation is discussed in section 3.1.2.

4.1.3 Possible Challenges/Barriers

During initial investigation into this topic, there were a few limitations that we encountered. Some of these, such as explainabiliy limitations, have lead to future research areas. I hope to explore many of these topics as a part of my broader dissertation.

4.1.3.1 Satellite Information

The satellite imagery we incorporated into this study had a number of notable limitations. First, while a satellite scene might contain 50% or less cloud cover (see figure 4.2), the clipped images might be completely covered in clouds (see, for example, figure 4.3). Further, in some cases the conflict event selected may be at the edge of a scene, with no valid scene available to fill in null information, resulting in a partially clipped image (see figure 4.1). Additionally, some of the clips contain interference or distortion, such as the clip at the bottom of figure 4.1.

Inter-related with these challenges, in many scenes we were unable to identify enough geographic locations to support the creation of 10 null cases. For example, in figure 4.4 we can see that the riot location in consideration does not have any null location possibilities due to the riot's proximity to the coast, and the concomitant lack of proximate urban areas eligible for building null (no-protest) cases. There are similar limitations that cause the distribution of clipped images in figure 3.10.

Another limitation is in our definition of where conflict events occurred, as the definition of a "neighborhood" is inherently imprecise. We used OpenStreetMaps (OpenStreetMap Contributors, 2024) to visually compare the size of our ten most repeated locations 3.3. We were able to confirm that the sizes of neighborhoods were inconsistent, but rarely of a size greater than our 10 square kilometer exclusionary zone (see figure 3.3).

To overcome this limitation, a future research direction might include ways to filter out clipped images with clouds or distortion. Additionally if we only exclude clipped images within 1 km of the edge of the satellite scene, we could

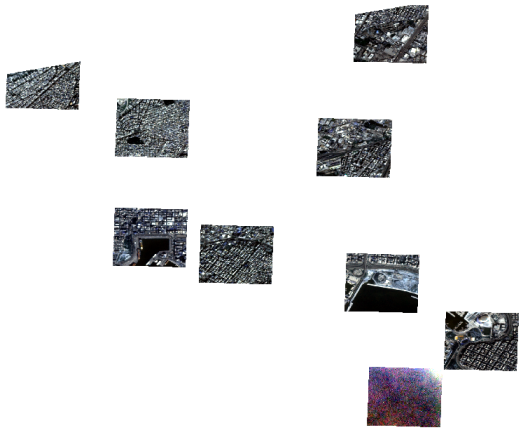


FIGURE 4.1: 9 of the null riot clipped images from Athens, Greece. Imagery ©Planet Labs PBC 2023. All rights reserved.

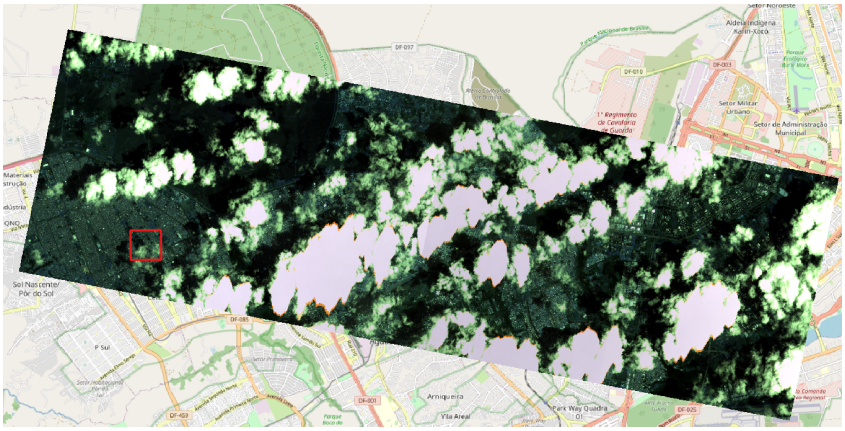


FIGURE 4.2: Satellite image of Brazil collected on 1 November 2018. This image contains less than 50% cloud cover for the full satellite scene. The riot location indicated in the red square has minimal cloud cover, but other locations in the scene will be impacted by the cloud cover as seen in figure 4.3. Imagery ©Planet Labs PBC 2023. All rights reserved.

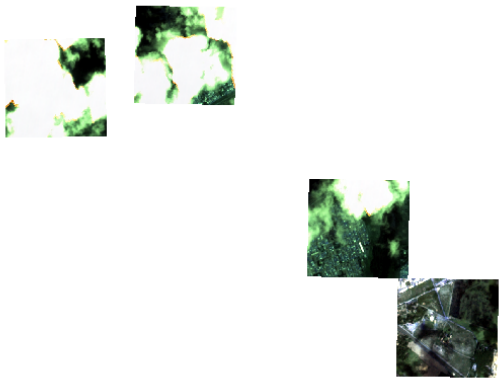


FIGURE 4.3: Clips from a satellite image of Bazil collected on 1 November 2018. While the full image contains less than 50% cloud cover, many of the clips are partially or completely obscured. Imagery ©Planet Labs PBC 2023. All rights reserved.

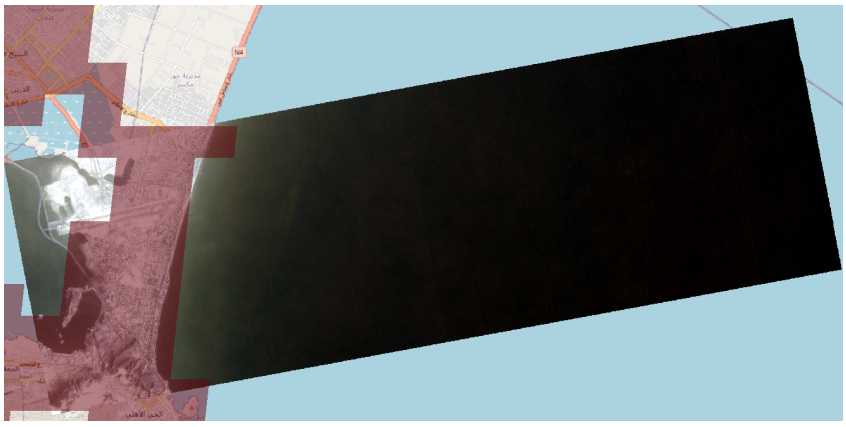


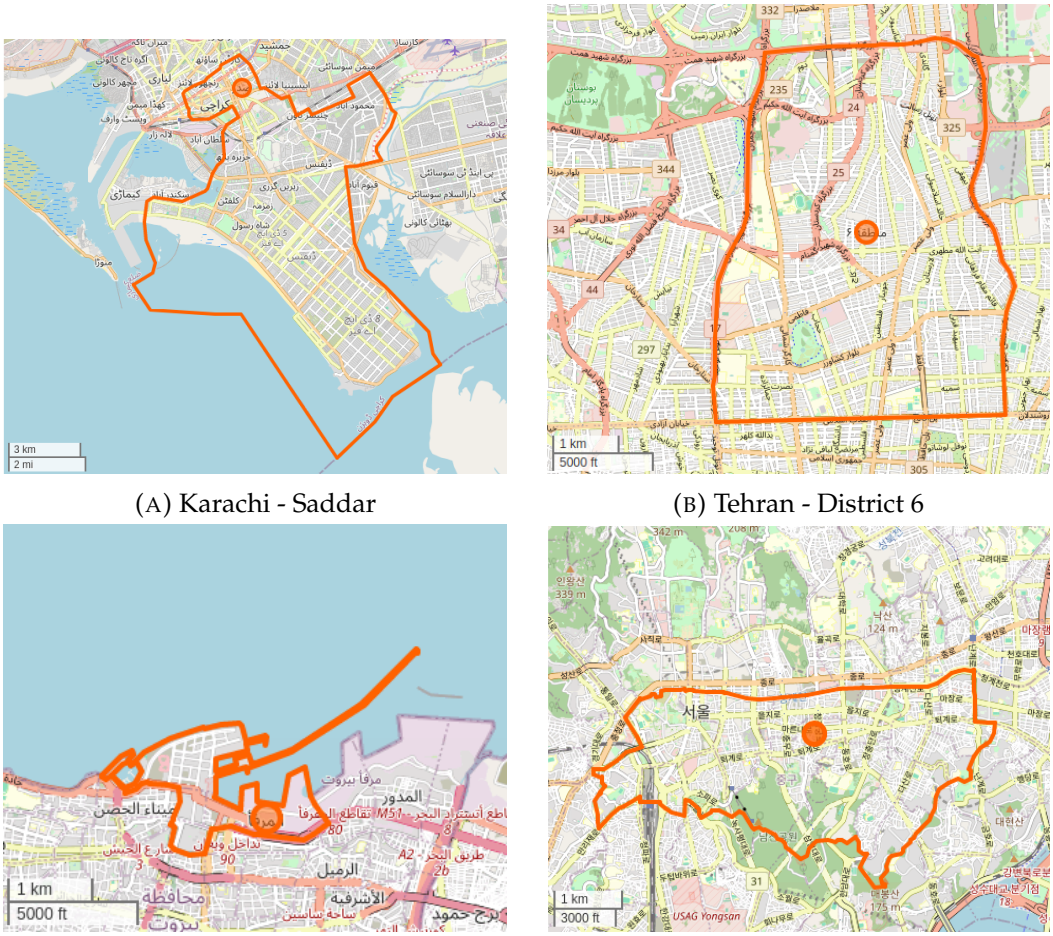
FIGURE 4.4: Satellite Image from Yemen collected on 14 September 2020. The urban areas are shown in red. Most of this image is not usable because of the lack of urban areas. Imagery ©Planet Labs PBC 2023. All rights reserved.

avoid images with only partial satellite information. The neighborhood limitation is more difficult to account for, but we might group locations by their average neighborhood size and train the groups individually in an attempt to train our network to learn patterns that account for neighborhood size differences.

4.1.3.2 Explainability

Currently, the majority of explainability techniques in the literature are focused on datasets consisting of object-centric images. For example, two common data sets CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton, et al., 2009) are used in many computer vision tasks and competitions, but those data sets only have objects centered in the middle of the picture, taking up most of the image space. This differs significantly from our satellite imagery. Our images contain all of the spatial information within a square kilometer in a city. As opposed to an image of a cat or dog, our images have multiple buildings, cars, streets, parks, etc. So while current explainability techniques can highlight portions of our image that lead to classification which are easily human interpretable, it is challenging for us to determine what in the image is being highlighted. The example we discussed previously, sports stadiums, are identified in Score-CAM and easily identified visually in the satellite image. There were other patterns that emerged in our Score-CAM analysis; however it is very difficult to describe many of the features Score-CAM identifies with easily identifiable semantic definitions. While we were able to identify a few other patterns, such as transitions from one zone to another zone (residential to commercial as an example), we are not confident in interpreting what these different types of zones are at this time. The field of explainability, as it relates to satellite images, has very little published in literature and remains a strong avenue for future inquiry.

To overcome this limitation, a future research direction is outlined in section 4.2. We intend to explore better understanding objects and spatial features identified by Score-CAM, through the use of points of interest identified in open source databases such as Overture Maps (Overture Maps Foundation, 2024). This would require segmenting the results of Score-CAM and comparing the objects indicated in Overture, to determine if there are patterns that arise indicating a riot or non-riot.



(A) Karachi - Saddar

(B) Tehran - District 6

(C) Beriut - Port

(D) Seoul - Jung

FIGURE 4.5: Four example neighborhoods from the top ten repeated locations. The scale of the image is given in the lower left corner of each image. The size of the neighborhoods is not consistent across the globe, but based of the labels in OpenStreetMap(OpenStreetMap Contributors, 2024), our methodology of excluding 10-km around the neighborhood will force our null cases to generate from locations outside the given neighborhood.

4.1.3.3 Additional Limitations

There are a number of additional limitations of the presented work. First, our data is focused on spatial information, not temporal, and thus we do not generate predictions of *when* a riot will occur, only the likely urban locations. Leveraging changes in images over time could help us overcome this challenge, but will necessitate new modeling strategies beyond those presented in this piece. Second, we have selected a ResNet18 as our base model, which could limit our model performance if alternative architectures are better performing.

Third, the ACLED database used to construct our imagery data set is drawn primarily from news sources (*ACLED Codebook* 2023). These comes with some inherent challenges and limitations. If riots and protests are occurring in regions that traditional news sources are not reporting about, the events are not likely to populate the ACLED database. Further, the nature of civil unrest is sometimes difficult to delineate with clear definitions, and different news organizations may cover a protest in conflicting ways - for example, a protest that is met with armed government resistance (*ACLED Codebook* 2023). These challenges are not likely to be overcome in the near term, but are notable as they may impact the results presented in this study.

To overcome this limitation, a future research direction might include downloading new images from weeks or months before the riot, as well as images from after the riot date. This would provide temporal differences for us to train our model to learn to differentiate. We could do this with a ResNet, or attempt to train with alternate architectures. To account for biases that might be present in the ACLED database, we might investigate alternate events within ALCED, or different databases.

4.2 Dissertation Paper 2: Explainability in Satellite Imagery

While initial results from my first dissertation paper indicate that satellite imagery can be used to predict - with up to 97% accuracy - where riots and protests are likely to occur in an urban area, it provides no insights into *what features in the image are important*. This is a critical step, as for policymakers to learn from

or trust this model, we must understand the key on-the-ground factors that the model is leveraging. Thus, while my first chapter focused on the estimation of the location of a conflict event, in this chapter I will focus on explaining the features that were actually used.

4.2.1 Major Research Question

Can we semantically describe the features within a satellite image that are consequential to the classification of urban conflict?

4.2.2 Data

The core satellite imagery data used for this analysis will be the same dataset described in section 4.1.2.1. This dataset provides over 2,178 images, which are each labeled as either a location at which protests or riots occurred, or an area drawn from the same urban environment at which no event occurred (see figure 3.3). Each of these images is a 1 km^2 box centered around a known riot/protest, or a non-riot location from the same satellite image as the riot clip. As a part of the quantitative analysis presented in Chapter 3, this data has already been acquired and processed.

In addition to satellite imagery, this paper's activities will require a method through which the features that are identified as important on the ground are semantically labeled. For example, a series of pixels that appears to represent a football stadium may be highlighted; in order to programmatically identify that location as a football stadium, we require a broad source of place names that we can correlate with the highlighted location. To accommodate this, we will leverage data sources from open source repositories, including SpaceNet(SpaceNet, 2024), DeepGlobe(Deep Globe, 2024), xView(Defense Innovation Unit, 2024)), Open Street Map (OpenStreetMap Contributors, 2024), and/or Overture Maps (Overture Maps Foundation, 2024). All of these are open source and freely available, but at this time it is unknown which of these - or which combination of these - will be best suited for our purposes.

4.2.3 Methods

Our core aim is, given an image of a location and a deep learning model designed to estimate if conflict is likely to occur at that location or not, provide a human-understandable, semantic description of why the model made a given decision. An output of this approach may be a sentence similar to one of these examples:
"It is likely that a riot will occur in this area, as it is proximate to numerous pubs and large public gathering spaces, in an area with a historic proclivity for riots."
"It is unlikely that a riot will occur in this area, as there is a lack of large physical spaces in which crowds could aggregate."
"We estimate, with 85% confidence, that a riot is likely to occur in this area. The key indicators included the presence of government properties and large fields in which aggregations can occur."

In order to accomplish this, we propose a multiple-step process, in which:

1. We implement a modified Score-CAM (SAT-Cam) to capture the pixels which were important to the model's decision function.
2. We threshold and then identify place names near the most important areas.
3. We validate based on consensus between data sources and human labeled observations.

4.2.3.1 Step 1. Modified Score-CAM

To date, explainability techniques in computer vision have predominantly emerged from traditional application - such as seeking to understand why an image taken from a cellphone is a cat or dog. As noted in the introduction to this prospectus, this has resulted in a number of limitations when it comes to applications for satellite imagery-based analyses. Critical themes include: **1.** Many current explainability techniques focus on images with a single object in the scene, satellite imagery by nature has many objects present in the same scene. **2.** Satellite imagery will have less variance in pixel values when compared to other images used in computer vision tasks, which will make understanding which spatial features are important in classification challenging to decipher. **3.** Semantic definitions are unclear in many satellite images, which will require us to use third-party data sources to distinguish features identified in explainability techniques.

In order to overcome these limits, we propose a modified implementation of Score-CAM. Score CAM is a class activation map explainability method that utilizes weights as opposed to gradient in the model. This method has demonstrated the ability to handle multi-target images better than other CAM methods (Wang et al., 2020). It is formally defined as:

$$L^c_{Score-CAM} = ReLU(\sum_k \alpha^c_k A^k_l) \tag{4.1}$$

A^k_l is the activation map for k^{th} channel of layer l, and α^c_k is the weight associated with the k^{th} neuron. Where $\alpha^c_k = C(A^k_l)$ for convolutional layer l and a class of interest c. And $C(A^k_l)$ is the channel-wise increase of confidence introduced with Score-CAM. (Wang et al., 2020).

In order to overcome issues associated with multi-target limits and variance, we propose two changes to Score-CAM. This revised algorithm, which we label "SAT-CAM", takes the following functional form:

$$L^c_{\text{SAT-CAM}} = \text{ReLU}\left(\sum_k \begin{cases} \alpha^k_c A^l_k & \text{if } \alpha^k_c A^l_k > \epsilon \\ 0 & \text{otherwise} \end{cases}\right) \tag{4.2}$$

Where $L^c_{SAT-CAM}$ is the output of SAT-CAM, that thresholds the Score-CAM evaluation $ReLU(\sum_k \alpha^c_k A^k_l)$ by ϵ. In effect, this removes pixels below the threshold ϵ for consideration.

The output of SAT-Cam is a heat map matrix of the activation layer A^k_l in which highlighted pixels contributed to the decision function. Building on Score-CAM, A^k_l is the last convolutional layer before the first fully connected layer. This is passed into the next stage of the process, in which we seek to use ancillary geographic databases to semantically label the regions of the image that are identified as important.

4.2.3.2 Step 2. Threshold & Identifying Place-Names

Using the output generated from SAT-CAM in equation 4.2, we will apply a segmentation algorithm (Kirillov et al., 2023), and then select the top 1 to 3 segments in terms of their absolute average value of importance. Once selected, we will test three different approaches to semantic interpretation (two stand-alone approaches, and one integrated). These are:

1. Clipping the underlying satellite imagery and passing it into an existing classification model.

2. Identifying the place-names that are proximate to important areas from ancillary geospatial databases, and integrating them into a human-readable form.

3. Providing both types of information into a large language model (LLM) to create a human-interpretable semantic description.

In the first approach, we seek to have a separate computer vision model label the isolated regions with their predominant structural characteristics. For example, there are a number of challenges, SpaceNet Challenge (SpaceNet, 2024), DeepGlobe Challenge (Deep Globe, 2024), and xView Detection Challenge (Defense Innovation Unit, 2024), that identify and label objects in satellite imagery. This will leverage existing models, such as SpaceNet to decipher what objects are in the imagery. If we apply these models to our regions of interest, we can generate a list of objects that appear frequently in our data. By analyzing the results we might be able to determine what objects or spatial characteristics and features are critical in classification.

In the second approach, we will leverage data from sources such as Overture Maps (Overture Maps Foundation, 2024) and Open Street Map (OpenStreetMap Contributors, 2024) to identify important place names proximate to the important features in geographic space. First this will require construction of segments that contain the regions of interest identified in SAT-CAM. These segments will have to be properly geo-referenced, so we only capture points of interest relevant to classification. Using these segments as limiting constraints, we will then generate a list of the points on interest from Overture, as an example.

These databases contain many types of information that might prove useful in explaining classification. Overture Maps Foundation (Overture Maps Foundation, 2024), for example, contains over 59 million locations. Each location in the Overture map data contains not only location information (latitude and longitude), but also some class information, such as "residential", "commercial", or "education." There are many land use classes as well, with examples such as "agriculture", "airport", "park", "recreation", "religious", and many more.

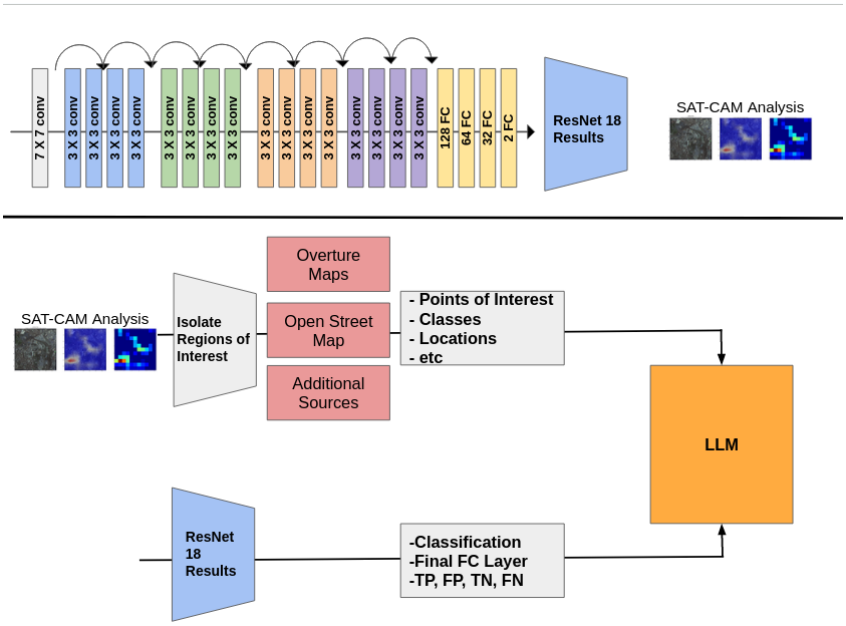


FIGURE 4.6: Flow chart detailing inputs into LLM. We will use the outputs of our ResNet18 and SAT-CAM analysis to feed text into an LLM.

This approach generates a list of location names, classes, and land uses for every segment identified by SAT-CAM. Additionally, we have the classification of the overall image from the ResNet (riot or non-riot), the output of the final fully connected layer, and the accuracy of the classification (True Positive, False Negative, etc). There is the potential to then leverage a LLM (Large Language Model) to help describe what is important in explaining classification. A diagram displaying the data generation that will be used as input into the LLM is shown in figure 4.6. At this point, all of the inputs would be text, and a trained LLM may be able to provide useful semantic descriptions given both the satellite-derived and gazetteer-based inputs.

A third approach, we will use a hybrid between the first two, in which a LLM will receive an input a prompt that contains all of the text input highlighted in the second approach (see figure 4.6) with the classification model utilized in the first approach. This approach is more complicated, but leverages both the visual classification models that already exist and additional information about the segmented regions available in open source database.

4.2.3.3 Step 3. Validation

There are two potential courses of action for validation of SAT-CAM, human labeling and consensus of databases. The first viable option is to simply have individuals label by hand the relevant regions of interest and compare the human labels with the SAT-CAM labels. This would be an relatively straightforward method when there is a single region of interest and a single human generated label. However, as the number of regions of interest or the size of the regions increase, there will be an associated increase in the amount of data humans are required to label. This increase spatial area has an accompanying increase in challenge and ambiguity. For example, if half of a clipped image is categorized as important by SAT-CAM, half a square kilometer of an urban area must be labeled by a human. Human labeling might prove easier to implement for smaller SAT-CAM results, and difficult to implement for large SAT-CAM results.

The second viable option for validation could be consensus. The results of the SAT-CAM analysis require access to databases with labeled locations. Once the methodology and data pipelines are established to evaluate a clipped image with SAT-CAM and generate a list of points of interest that exist within the segmented satellite data, the process can be replicated with different open source data bases. If the list of locations for Overture Maps, Open Street Map, and potentially others, all agree, then the list is most likely an accurate representation of the spatial features important for classification.

4.2.4 Possible Challenges/Barriers

There are a number of possible challenges to the successful execution of this work. First and foremost is a lack of relevant past literature on which to build. The majority of prior work in literature pertaining to computer vision uses data sets and examples that are object-centric. Much of the prior work uses standard data sets such as CIFAR-10 and CIFAR-100, which consist exclusively of photographic images, with the object centered in frame. Satellite imagery is very different in practice. The dataset under examination in this study comprises satellite imagery, distinct from conventional photographs, encompassing spatial regions of interest that may appear at any location within the image frame,

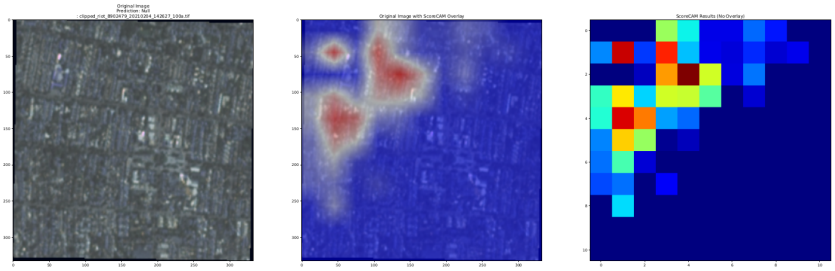


FIGURE 4.7: Score-CAM results displayed on the right of this figure represent a single contiguous region, regardless of the threshold set for consideration into a region of interest.

rather than being centrally positioned. This characteristic marks a significant departure from the datasets predominantly utilized in existing literature.

Another potential challenge is that the Score-CAM results will contain both contiguous and non-contiguous results. If all of the regions of interest are contiguous (as in figure 4.7), the association of the single region as the single entity that is important to classification is straight forward. However, there are other Score-CAM results that are not contiguous (see figure 4.8), and it is unclear if these regions should be evaluated individually or collectively.

Determining how to treat these non-contiguous areas as a non-overlapping polygons in the explainability research is similar to the object based image analysis (OBIA) challenge of analyzing images and objects at the appropriate resolution to avoid the "salt and pepper effect" (Blaschke, 2010). So while there is similar work with satellite imagery, in fields such as OBIA, there is little work exploring the potential obstacles associated with explainability. The non-contiguous nature of important features in satellite imagery could cause a number of challenges, opening a number of areas for inquiry. Questions include: Is the classification determined by the sum of all weights or regions, or the maximum weight of a single region? What role do interaction effects of non-contiguous regions play in classification? Does the spatial distance between non-contiguous regions impact the classification of the image? These challenges are specific to the use of explainability techniques in satellite images.

Integral to the previously challenges is the underlying decision in determining what threshold to establish for consideration as a region of interest. Before

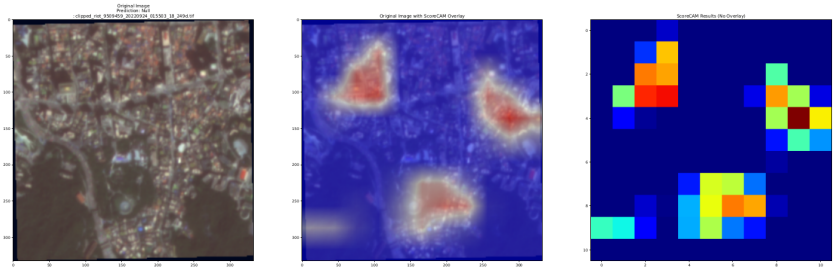


FIGURE 4.8: Score-CAM results displayed on the right of this figure represent a multiple regions of interest, regardless of the threshold set for consideration into a region of interest.

constructing the regions of interest, we must determine how much of the Score-CAM results should be used to consider. If we set the threshold (ϵ in equation 4.2) too low, we will consider too much of image. If we consider too much of the image, the task of determining what in the image is being identified can be overwhelmed with too much non-essential information. If we set the threshold ϵ too high, we will not consider enough of the image. This will eliminate too much of the image, making identification, in terms of human interpretation, very challenging. If we refer back to the sports stadium example discussed in Chapter 3, if we eliminate too much of the stadium because of a high threshold, we might only see some of the stadium making identification more challenging.

4.3 Dissertation Paper 3: Identification of conflict within Satellite Imagery

4.3.1 Major Research Question

Can deep learning techniques enable the localization of conflict across a full-sized satellite image?

While the first paper of my dissertation examines *if* we can predict the likelihood of conflict using paired cases of images, and paper two explores *why* these predictions are made, this leaves the question of if these tools can be operationally useful open. While an algorithm that can discriminate between pairs of conflict and no-conflict cases may be useful, in practice it is unlikely that practitioners will have such a database. More practically, they are likely to have many

satellite scenes that they would like to identify probable locations of riots or protests. While a model designed to take in pairs of images can accommodate this task, it requires additional modeling and testing in that broader context. This chapter will fill that gap.

4.3.1.1 Data

My third paper will continue to leverage information from Planet, focusing on full images from the satellite sensors rather than clipped regions. We plan to use the same source of data as specified in section 4.1.2.1, which consists of 18,631 full satellite scenes distributed around the world. These images were collected from October 2017 to September 2022 and selected to be at least 24 hours prior to a conflict event. An example of a single image collected from a satellite can be seen in figure 3.1. We also maintain a record of the latitude and longitude, as well as the clipped image, of the known riot for each full satellite scene.

4.3.1.2 Methods

We will use our trained ResNet18 from chapter 3 to identify the likely locations of riots and protests in full satellite scenes. To accomplish this, we will take the following steps:

1. Create a grid that subsets the full satellite scene.
2. Systematically evaluate 1 km^2 boxes, encompassing the full satellite scene.
3. Collect results from step 2, to identify locations of probable riot or protest.

In the first step of our methodology, we establish a grid that subsets the entire satellite scene into half kilometer boxes (see figure 4.9). We subset the satellite scene, because we aim to evaluate the scene in a similar manner to a convolutional filter passing over a full image. Our goal is to evaluate groups of subset half kilometer boxes in the same way a 2x2 convolutional filter captures the spatial relationship of four adjacent pixels in an input image, resulting in a single output value for the adjacent pixels. We will evaluate four of the subset grids as a single clipped image and generate a single output, in our scenario a classification of riot or non-riot. By subseting the full satellite scene into half kilometer

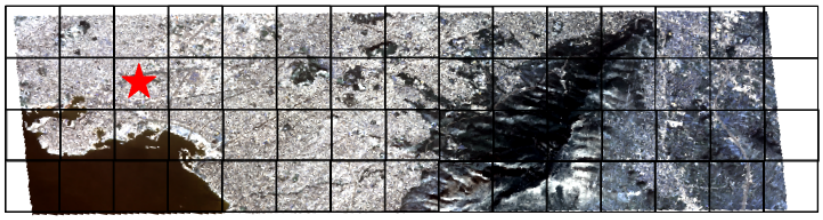


FIGURE 4.9: Example of grid that subsets entire satellite scene. This image is not to scale, but represents a potential half kilometer grid that encompasses the full satellite scene. In this example, there is a known feature(s) that causes a riot classification represented by a red star. Imagery ©Planet Labs PBC 2023. All rights reserved.

boxes, we preserve the capability of our trained ResNet18 to evaluate each 2x2 grouping as an individual clipped image, similar to our data set in chapter 3. This subseting of full satellite scenes will be unique to each individual scene. As discussed in chapter 3, and highlighted in table 3.2, the spatial dimensions of images vary with each generation of satellite as well as the off-nadir angle from when the the image is captured.

In the second step of our methodology, we evaluate the 2x2 adjacent half kilometer boxes across the full satellite scene. Again, we build off the analogy of a convolutional filter sliding across an input image. We will pad our satellite scene to account for data at the edge of the scene. We implement a stride of one, to evaluate each half kilometer box multiple times. An illustration of this process is shown in figure 4.10. The purpose of subsetting the image and evaluating each half kilometer box as part of a 2x2 grouping, is to account for the unknown proximity relationship among spatial features that are important to riot classification. This technique allows us to evaluate spatial features across a wider area of consideration, as opposed to limiting the spatial features to evaluation in the isolation of a single kilometer or half kilometer box.

The third step is to collect the results of the second step to identify the probable locations for protests or riots. Continuing with the 2x2 convolutional filter analogy, when the ResNet18 evaluates the four adjacent half kilometer boxes, all four are categorized as a riot or non-riot. Each individual half kilometer box is evaluated four times, with different combinations of the surrounding boxes. If the spatial features important to riot classification are present in a particular box, then that half kilometer box should be categorized as a riot four times. While

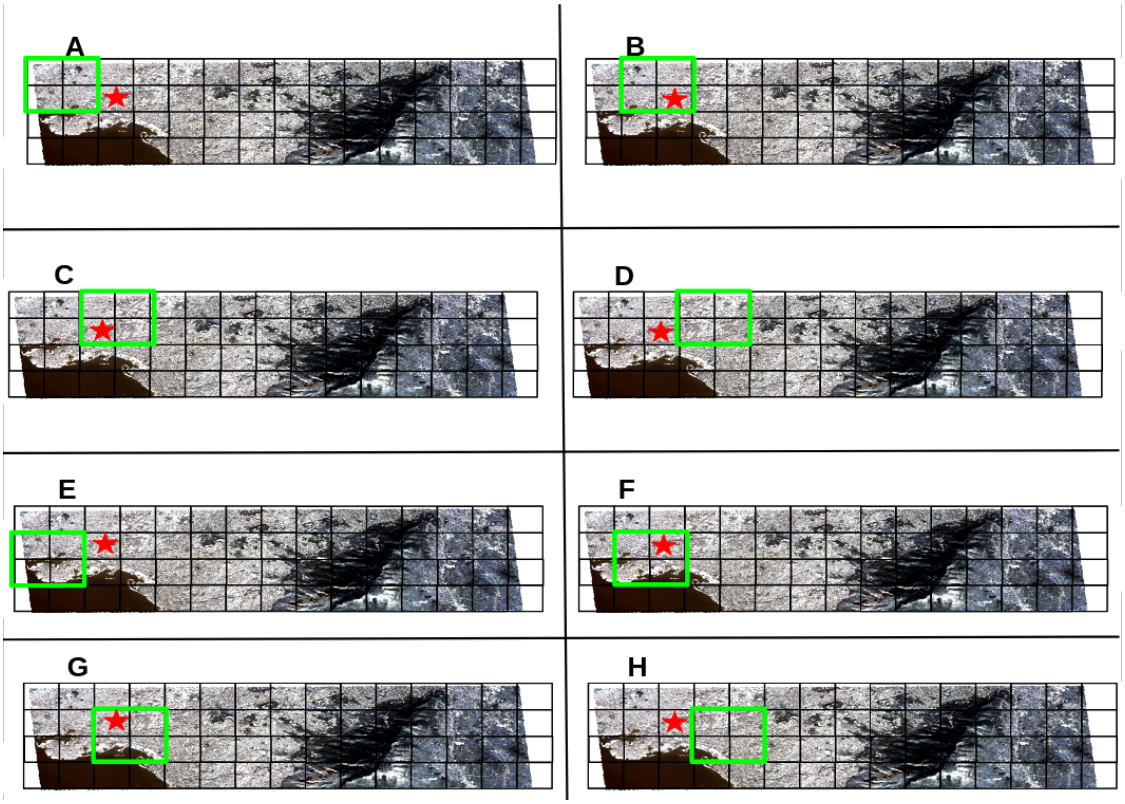


FIGURE 4.10: This illustration is not to scale. The green box represents the 2x2 convolutional filter used to evaluate the satellite scene. Again, there is a known feature(s) that causes a riot classification represented by a red star. The green box slides across the full scene. In frames B, C, F, and G, all four of the half kilometer boxes would be marked as a riot. In frames A, D, E, and H, none of the half kilometer boxes would be marked as a riot. Imagery ©Planet Labs PBC 2023. All rights reserved.

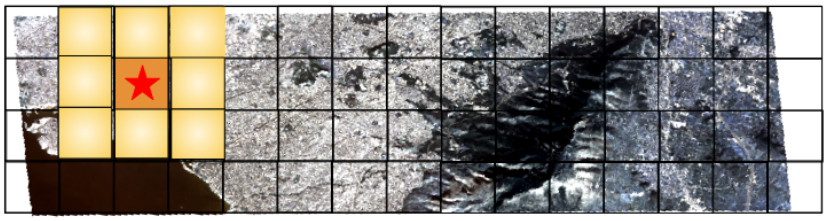


FIGURE 4.11: The resulting heat map created during evaluation, that contains the spatial features that cause riot classification indicated by the red star. Since the half kilometer box containing the red star was counted as a riot more often than its neighbors, this region appears as a hot spot that indicated probable riot or protest. Illustration is not to scale. Imagery ©Planet Labs PBC 2023. All rights reserved.

adjacent half kilometer boxes are categorized as a riot less than four times. This will create a heat map across the full satellite scene, highlighting the particular regions in the image that are probable locations for riot or protest (see figure 4.11). We then compare the heat map indicated locations to the known riot location to determine if the ResNet is able to determine the location of the riot or protest from a full satellite scene.

There are two ways to validate our findings in step three. The first is to measure the distance from the center of the region indicated in the heat map, to the latitude and longitude of the known riot. The second is to compare how much the clipped image overlaps with the hot spots from the heat map. It is currently unknown if our outlined methodology will result in the identification of a single riot or multiple riots. We also do not know the size of any of the hot spots our methodology will create. We are currently exploring the most appropriate manner for evaluation of performance in terms of accuracy, precision, recall, or other relevant measures.

There is potential to include further analysis related to explainability highlighted in section 4.2. After identifying regions in the full image that are likely to have a protest, we can implement SAT-CAM to gain insight into the features driving classification. It is currently unknown how the relationship between spatial features and their distribution impact classification. Additionally, we do not know how spatial separation among important features influences classification across wider areas, such as full satellite scenes. For example, large parks

and the presence of government buildings might drive riot classification in a single one kilometer clipped image, but those same features when separated by a greater distance might no longer indicate a riot.

4.3.2 Possible Challenges/Barriers

There are a few potential challenges with this methodology. One of the key assumptions in our training data is that there is only one riot or protest in each satellite scene, or at least there are no riots outside the ten kilometer exclusion area when the null clips were created (as described in chapter 3). If there are other riots in the satellite scene, the trained network might correctly identify riot locations outside the known clip location, and based on our training data we would consider that a miss-classification.

Another potential challenge identified for this study involves the methodology of analyzing entire satellite scenes to detect instances of riots or protests using a network originally trained on one-square-kilometer segments. The underlying assumption of our model posits that the spatial characteristics relevant to the classification task are less than one kilometer in size and that any spatial correlations critical for accurate classification occur within this distance. However, this assumption may not hold in all cases, as it is conceivable that certain spatial features or indicators of a riot or protest exceed this size limit or are dispersed beyond a kilometer, leading to potential inaccuracies in classification. This challenge underscores the need for a robust evaluation of the model's assumptions and its capability to generalize across varying spatial dimensions.

A further potential challenge originates from parameter choice in our methodology. We opt for a half kilometer segmentation in order to evaluate each half kilometer box multiple times with its adjacent neighbors. While this allows each segment to be classified multiple times, we might not be isolating the relevant spatial features enough. We could attempt to compensate for this by decreasing the size of our grid. For example we could use a quarter kilometer grid, and include a filter with a 4 x 4 block to construct a kilometer box for evaluation. This would allow us to evaluate the relevant spatial feature more often, but it would also increase the computational cost. We do not know the size of the features that are relevant to prediction, nor do we know if they are consistent in size.

So we are presented with an unknown constraint to the appropriate size of the grid to best capture the relevant features. This problem is similar to the issue of required spatial resolution for object based image analysis (Blaschke, 2010)

Chapter 5

Timeline for Degree Completion

5.1 Gaant Chart

The work described in my quantitative section, chapter 3 and section 4.1 was submitted to *Transactions in Geographic Information Science* on 15 March 2024, and is currently under review (shown in blue in figure 5.1). This is important in subsequent phases of my dissertation. This initial research is responsible for generating a significant portion of the data required for follow on phases. This included the selection, download, and pre-processing of over 4 TB of satellite imagery data. The amount of work invested in constructing this data set, should facilitate the ambitious timeline I present to support a May 2025 graduation (see figure 5.1).

Specifically, the first paper included preliminary work into explainability in satellite imagery. The goal of the second phase of the research (purple in figure 5.1) is to identify specific features in the image that might be important in classification of the image as either a riot or not. The first two tasks in this are to gather point of interest data from Overture Maps (Overture Maps Foundation, 2024), and to segment the pixels of interest from the Score-CAM analysis. These two tasks can be accomplished concurrently during the month of April, given current progress. During May, I will work on an algorithm which integrates points of interest with the pixels of interest to determine patterns of physical features on the ground that the trained ResNet has learned to associate with riots or null riots. A goal of completing this phase by the summer of 2024, should be attainable and provide a much richer understanding of explainability as it pertains to satellite imagery.

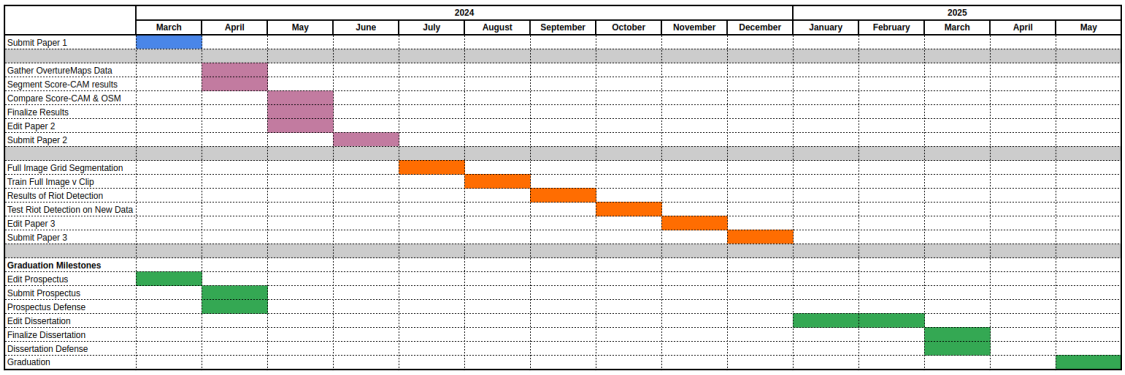


FIGURE 5.1: Gantt Chart supporting a May 2025 graduation.

The third chapter of my dissertation (orange in figure 5.1) involves identifying conflict from contiguous satellite images, contrasted with classification of smaller 1 km^2 boxes in the first phase. The first task in this phase is to determine the appropriate methodology to segment full satellite images into a grid for subsequent conflict searches. This effort should begin in July 2024. By August, I will be training neural networks to detect riots within these full satellite scenes. This will likely result in refinement in either the grid segmentation of satellite scenes, or the training of neural networks. Depending on the initial results of this, there is a potential that I will have to increase the volume of my current data set or alter my methodology. My current plan is to be complete with this research effort by the end of 2024.

I propose the following anticipated timeline, with consideration given to the dissertation milestones (indicated in green in Figure 5.1), which are aligned to support a graduation in May 2025. In my initial paper, I aim to contribute to the body of literature on deep learning, conflict, and the classification of satellite imagery in urban environments. Subsequently, my second paper will delve into novel areas of explainability within the context of satellite imagery. Lastly, in my third paper, I plan to broaden the scope of my research by extending the functionality from analyzing small urban areas to encompassing comprehensive satellite scenes of larger areas.

Bibliography

ACLED (2022). *ACLED:Bringing clarity to crisis*. URL: https://acleddata.com/ (visited on 11/03/2022).

ACLED Codebook (2023). Armed Conflict Location & Event Data Project (ACLED). URL: www.acleddata.com.

Alo, Clement Aga and Robert Gilmore Pontius Jr (2008). "Identifying systematic land-cover transitions using remote sensing and GIS: the fate of forests inside and outside protected areas of Southwestern Ghana". In: *Environment and Planning B: Planning and Design* 35.2, pp. 280–295.

Alsaedi, Nasser, Pete Burnap, and Omer Rana (2015). "Identifying disruptive events from social media to enhance situational awareness". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 934–941.

— (2017). "Can we predict a riot? Disruptive event detection using Twitter". In: *ACM Transactions on Internet Technology (TOIT)* 17.2, pp. 1–26.

Anderson, Christopher J and Silvia M Mendes (2006). "Learning to lose: Election outcomes, democratic experience and political protest potential". In: *British Journal of Political Science* 36.1, pp. 91–111.

Andronikidou, Aikaterini and Iosif Kovras (2012). "Cultures of rioting and anti-systemic politics in Southern Europe". In: *West European Politics* 35.4, pp. 707–725.

Aung, Thiri Shwesin et al. (2021). "Using satellite data and machine learning to study conflict-induced environmental and socioeconomic destruction in data-poor conflict areas: The case of the Rakhine conflict". In: *Environmental Research Communications* 3.2, p. 025005.

Babenko, Boris et al. (2017). "Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico". In: *arXiv preprint arXiv:1711.06323*.

Becker, Hila, Mor Naaman, and Luis Gravano (2011). "Beyond trending topics: Real-world event identification on twitter". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 5. 1, pp. 438–441.

Begum, Salma (2018). "Changing Scenarios of Public Open Space in a British Colonial City: The Case of the Ramna Area, Dhaka". In: *Nakhara: Journal of Environmental Design and Planning* 14, pp. 39–56.

Bencsik, Panka (2018). "The non-financial costs of violent public disturbances: Emotional responses to the 2011 riots in England". In: *Journal of Housing Economics* 40, pp. 73–82.

Berazneva, Julia and David R Lee (2013). "Explaining the African food riots of 2007–2008: An empirical analysis". In: *Food policy* 39, pp. 28–39.

Bharti, Nita and Andrew J Tatem (2018). "Fluctuations in anthropogenic nighttime lights from satellite imagery for five cities in Niger and Nigeria". In: *Scientific data* 5.1, pp. 1–9.

Bibault, Jean-Emmanuel et al. (2020). "Deep learning prediction of cancer prevalence from satellite imagery". In: *Cancers* 12.12, p. 3844.

Blaschke, Thomas (2010). "Object based image analysis for remote sensing". In: *ISPRS journal of photogrammetry and remote sensing* 65.1, pp. 2–16.

Bonnasse-Gahot, Laurent et al. (2018). "Epidemiological modelling of the 2005 French riots: a spreading wave and the role of contagion". In: *Scientific reports* 8.1, p. 107.

Brewer, Ethan, Jason Lin, and Dan Runfola (2022). "Susceptibility & defense of satellite image-trained convolutional networks to backdoor attacks". In: *Information Sciences* 603, pp. 244–261.

Brewer, Ethan, Zhonghui Lv, and Dan Runfola (2023). "Tracking the industrial growth of modern China with high-resolution panchromatic imagery: A sequential convolutional approach". In: *arXiv preprint arXiv:2301.09620*.

Brewer, Ethan et al. (2021). "Predicting road quality using high resolution satellite imagery: A transfer learning approach". In: *Plos one* 16.7, e0253370.

Buhrmester, Vanessa, David Münch, and Michael Arens (2021). "Analysis of explainers of black box deep neural networks for computer vision: A survey". In: *Machine Learning and Knowledge Extraction* 3.4, pp. 966–989.

Cadena, Jose et al. (2015). "Forecasting social unrest using activity cascades". In: *PloS one* 10.6, e0128879.

Carleer, AP, Olivier Debeir, and Eléonore Wolff (2005). "Assessment of very high spatial resolution satellite image segmentations". In: *Photogrammetric Engineering & Remote Sensing* 71.11, pp. 1285–1294.

Carranza-García, Manuel, Jorge García-Gutiérrez, and José C Riquelme (2019). "A framework for evaluating land use and land cover classification using convolutional neural networks". In: *Remote Sensing* 11.3, p. 274.

Charuchinda, P et al. (2019). "On the use of class activation map for land cover mapping". In: *2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, pp. 653–656.

Chattopadhay, Aditya et al. (2018). "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 839–847.

Chauhan, Rahul, Kamal Kumar Ghanshala, and RC Joshi (2018). "Convolutional neural network (CNN) for image detection and recognition". In: *2018 first international conference on secure cyber computing and communication (ICSCCC)*. IEEE, pp. 278–282.

Ciorciari, John D and Jessica Chen Weiss (2016). "Nationalist protests, government responses, and the risk of escalation in interstate disputes". In: *Security Studies* 25.3, pp. 546–583.

Compton, Ryan et al. (2013). "Detecting future social unrest in unprocessed twitter data:"emerging phenomena and big data"". In: *2013 IEEE International Conference on Intelligence and Security Informatics*. IEEE, pp. 56–60.

Dabkowski, Piotr and Yarin Gal (2017). "Real time image saliency for black box classifiers". In: *Advances in neural information processing systems* 30.

Daudt, Rodrigo Caye et al. (2018). "Urban change detection for multispectral earth observation using convolutional neural networks". In: *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. Ieee, pp. 2115–2118.

Davies, Toby P et al. (2013). "A mathematical model of the London riots and their policing". In: *Scientific reports* 3.1, p. 1303.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Deep Globe (Feb. 2024). *Home Page*. URL: deepglobe.org/challenge.html (visited on 02/16/2024).

Defense Innovation Unit (Feb. 2024). *Home Page*. URL: xviewdataset.org (visited on 02/16/2024).

Dosovitskiy, Alexey and Thomas Brox (2016). "Inverting visual representations with convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837.

Eckstein, Susan (2001). "Power and Popular Protest in Latin". In: *Power and popular protest: Latin American social movements*, p. 1.

El Warea, Mohamad et al. (2019). "Riots in beirut: description of the impact of a new type of mass casualty event on the emergency system in Lebanon". In: *Disaster medicine and public health preparedness* 13.5-6, pp. 849–852.

Elvidge, Christopher D et al. (2012). "The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data". In: *Social Geography* 7.1, pp. 23–35.

European Commission and Statistical Office of the European Union (2021). *Applying the Degree of Urbanisation — A methodological manual to define cities, towns and rural areas for international comparisons — 2021 edition*. ISBN 978-92-76-20306-3, doi:10.2785/706535. Publications Office of the European Union.

European Space Agency (2015a). *Sentinel 2 Satellite Imagery Technical Data Sheet*. https://sentinels.copernicus.eu/web/sentinel/home. [Online; accessed 23-February-2024]. URL: https://sentinels.copernicus.eu/documents/247904/1848117/Sentinel-2_Data_Products_and_Access.

— (2015b). *SPOT-6 Satellite Imagery Technical Data*. https://earth.esa.int/eogatewaye. [Online; accessed 23-February-2024]. URL: https://earth.esa.int/eogateway/missions/spot-6.

Filchenkov, Andrey A, Artur A Azarov, and Maxim V Abramov (2014). "What is more predictable in social media: Election outcome or protest action?" In: *Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia*, pp. 157–161.

Fong, Ruth C and Andrea Vedaldi (2017). "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437.

Foody, Giles M and Ajay Mathur (2004). "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification". In: *Remote Sensing of Environment* 93.1-2, pp. 107–117.

Fox, Sean and Andrew Bell (2016). "Urban geography and protest mobilization in Africa". In: *Political Geography* 53, pp. 54–64.

Fradkov, Alexander L (2020). "Early history of machine learning". In: *IFAC-PapersOnLine* 53.2, pp. 1385–1390.

Fu, Kun et al. (2019). "Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images". In: *Remote sensing* 11.5, p. 544.

Gehlbach, Scott (2010). "Reflections on Putin and the Media". In: *Post-Soviet Affairs* 26.1, pp. 77–87.

Gong, Qiang and Rajan Batta (2007). "Allocation and reallocation of ambulances to casualty clusters in a disaster relief operation". In: *Iie Transactions* 39.1, pp. 27–39.

González-Bailón, Sandra et al. (2011). "The dynamics of protest recruitment through an online network". In: *Scientific reports* 1.1, pp. 1–7.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Goodman, Seth, Ariel BenYishay, and Daniel Runfola (2021). "A convolutional neural network approach to predict non-permissive environments from moderate-resolution imagery". In: *Transactions in GIS* 25.2, pp. 674–691.

Gorban, Alexander N, Evgeny M Mirkes, and Ivan Y Tyukin (2020). "How deep should be the depth of convolutional neural networks: a backyard dog case study". In: *Cognitive Computation* 12, pp. 388–397.

Greer, Chris and Eugene McLaughlin (2010). "We predict a riot? Public order policing, new media environments and the rise of the citizen journalist". In: *The British Journal of Criminology* 50.6, pp. 1041–1059.

Han, Xiaobing et al. (2017). "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification". In: *Remote Sensing* 9.8, p. 848.

He, Chunyang et al. (2019a). "Detecting global urban expansion over the last three decades using a fully convolutional network". In: *Environmental Research Letters* 14.3, p. 034008.

He, Kaiming et al. (2016a). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

— (2016b). "Identity mappings in deep residual networks". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pp. 630–645.

He, Tong et al. (2019b). "Bag of tricks for image classification with convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 558–567.

Helber, Patrick et al. (2019). "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7, pp. 2217–2226.

Heryadi, Yaya et al. (2020). "The effect of resnet model as feature extractor network to performance of DeepLabV3 model for semantic satellite image segmentation". In: *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*. IEEE, pp. 74–77.

Heslin, Alison (2021). "Riots and resources: How food access affects collective violence". In: *Journal of Peace Research* 58.2, pp. 199–214.

Ho, Yaoshiang and Samuel Wookey (2019). "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling". In: *IEEE access* 8, pp. 4806–4813.

Hu, Jie, Li Shen, and Gang Sun (2018). "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Hu, Wenjie et al. (2019). "Mapping missing population in rural India: A deep learning approach with satellite imagery". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 353–359.

Huang, Gao et al. (2016). "Deep networks with stochastic depth". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pp. 646–661.

Huang, Gao et al. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Jean, Neal et al. (2016). "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301, pp. 790–794.

Joya, Angela (2011). "The Egyptian revolution: crisis of neoliberalism and the potential for democratic politics". In: *Review of African political economy* 38.129, pp. 367–386.

Khan, Riaz Ullah et al. (2018). "Evaluating the performance of resnet model based on image recognition". In: *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pp. 86–90.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kirillov, Alexander et al. (2023). "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026.

Klein, Graig R and Patrick M Regan (2018). "Dynamics of political protests". In: *International Organization* 72.2, pp. 485–521.

Korolov, Rostyslav et al. (2016). "On predicting social unrest using social media". In: *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, pp. 89–95.

Kramer, Herbert J and Arthur P Cracknell (2008). "An overview of small satellites in remote sensing". In: *International journal of remote Sensing* 29.15, pp. 4285–4337.

Krizhevsky, Alex, Geoffrey Hinton, et al. (2009). *Learning multiple layers of features from tiny images*. http://www.cs.toronto.edu/~kriz/cifar.html.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Kussul, Nataliia et al. (2017). "Deep learning classification of land cover and crop types using remote sensing data". In: *IEEE Geoscience and Remote Sensing Letters* 14.5, pp. 778–782.

Leclerc, Maxime et al. (2018). "Ship classification using deep learning techniques for maritime target tracking". In: *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, pp. 737–744.

Lodhi, Bilal and Jaewoo Kang (2019). "Multipath-DenseNet: A Supervised ensemble architecture of densely connected convolutional networks". In: *Information Sciences* 482, pp. 63–72.

Löwenheim, Oded (2007). "The responsibility to responsibilize: Foreign offices and the issuing of travel warnings". In: *International Political Sociology* 1.3, pp. 203–221.

Lv, Zhonghui et al. (2024). "Mapping the tidal marshes of coastal Virginia: a hierarchical transfer learning approach". In: *GIScience & Remote Sensing* 61.1, p. 2287291.

Mahendran, Aravindh and Andrea Vedaldi (2015). "Understanding deep image representations by inverting them". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196.

Maxar (2020a). *GeoEye-1 Satellite Imagery Technical Data Sheet*. https://www.maxar.com. [Online; accessed 23-February-2024]. URL: https://resources.maxar.com/data-sheets/geoeye-1.

— (2020b). *WorldView-3 Satellite Imagery Technical Data Sheet*. https://www.maxar.com. [Online; accessed 23-February-2024]. URL: https://resources.maxar.com/data-sheets/worldview-3.

Nabiee, Shima et al. (2022). "Hybrid U-Net: Semantic segmentation of high-resolution satellite images to detect war destruction". In: *Machine Learning with Applications* 9, p. 100381.

Naidu, Rakshit et al. (2020). "IS-CAM: Integrated Score-CAM for axiomatic-based explanations". In: *arXiv preprint arXiv:2010.03023*.

Najjar, Alameen, Shun'ichi Kaneko, and Yoshikazu Miyanaga (2018). "Crime mapping from satellite imagery via deep learning". In: *arXiv preprint arXiv:1812.06764*.

Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han (2015). "Learning deconvolution network for semantic segmentation". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528.

Nordhaus, William D (2001). "The progress of computing". In: *Available at SSRN 285168*.

OpenStreetMap Contributors (2024). *OpenStreetMap*. Accessed: 2024. URL: https://www.openstreetmap.org.

Overture Maps Foundation (Feb. 2024). *Home Page*. URL: https://www.overturemaps.org (visited on 02/12/2024).

Pain, Paromita and Ezequiel Korin (2021). "'Everything is dimming out, little by little:'examining self-censorship among Venezuelan journalists". In: *Communication Research and Practice* 7.1, pp. 71–88.

Patel, Krishna, Chintan Bhatt, and Pier Luigi Mazzeo (2022). "Deep learning-based automatic detection of ships: An experimental study using satellite images". In: *Journal of imaging* 8.7, p. 182.

Pearce, Tim, Alexandra Brintrup, and Jun Zhu (2021). "Understanding softmax confidence and uncertainty". In: *arXiv preprint arXiv:2106.04972*.

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2010). "Streaming first story detection with application to twitter". In: *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pp. 181–189.

Petsiuk, Vitali, Abir Das, and Kate Saenko (2018). "Rise: Randomized input sampling for explanation of black-box models". In: *arXiv preprint arXiv:1806.07421*.

Phillips, Lawrence et al. (2017). "Using social media to predict the future: a systematic literature review". In: *arXiv preprint arXiv:1706.06134*.

Planet (2022a). *Planet Application Program Interface: In Space for Life on Earth*. Planet. URL: https://api.planet.com.

— (2022b). *PlanetScope: Constellation and sensor overview*. Planet. URL: https://developers.planet.com/docs/data/planetscope/.

Pond, Philip and Jeff Lewis (2019). "Riots and Twitter: connective politics, social media and framing discourses in the digital public sphere". In: *Information, Communication & Society* 22.2, pp. 213–231.

Purbrick, Martin (2019). "A report of the 2019 Hong Kong protests". In: *Asian Affairs* 50.4, pp. 465–487.

Rahimi, Babak (2015). "Censorship and the Islamic Republic: Two modes of regulatory measures for media in Iran". In: *The Middle East Journal* 69.3, pp. 358–378.

Renaud, Molly et al. (2019). "Social network structure as a predictor of social behavior: the case of protest in the 2016 us presidential election". In: *Recent Developments in Data Science and Intelligent Analysis of Information: Proceedings of the XVIII International Conference on Data Science and Intelligent Analysis of Information, June 4–7, 2018, Kyiv, Ukraine*. Springer, pp. 267–278.

Rodrigues, Marco TA et al. (2010). "Automatic fish species classification based on robust feature extraction techniques and artificial immune systems". In: *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*. IEEE, pp. 1518–1525.

Rogan, John and DongMei Chen (2004). "Remote sensing technology for mapping and monitoring land-cover and land-use change". In: *Progress in planning* 61.4, pp. 301–325.

Rosenblatt, Frank (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Runfola, D, A Stefanidis, and H Baier (2022). "Using satellite data and deep learning to estimate educational outcomes in data-sparse environments". In: *Remote Sensing Letters* 13.1, pp. 87–97.

Runfola, Dan et al. (2024). "A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance". In: *International Journal of Geographical Information Science*, pp. 1–25.

Runfola, Daniel et al. (2020). "geoBoundaries: A global database of political administrative boundaries". In: *PLoS One* 15.4, e0231866.

Runfola, Daniel et al. (2022). "Deep learning fusion of satellite and social information to estimate human migratory flows". In: *Transactions in GIS* 26.6, pp. 2495–2518.

Rüttgers, Mario et al. (2019). "Prediction of a typhoon track using a generative adversarial network and satellite images". In: *Scientific reports* 9.1, p. 6057.

Sattarzadeh, Sam et al. (2021). "Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1775–1779.

Schiavina, M., M. Melchiorri, and M. Pesaresi (2023). *GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030)*. European Commission, Joint Research Centre (JRC). PID: http://data.europa.eu/89h/a0df7a6f-49de-46ea-9bde-563437a6e2ba, doi:10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA.

Sekiyama, Taro et al. (2018). "Profile-guided memory optimization for deep neural networks". In: *arXiv preprint arXiv:1804.10001*.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shakur, Tasleem, Ishrat Islam, and Javeria Masood (2010). "WHAT CULTURE, WHOSE SPACE AND WHICH TECHNOLOGY? THE CONTESTED TRANSFORMATION AND THE CHANGING HISTORIC BUILT ENVIRONMENTS OF SOUTH ASIA." In: *ArchNet-IJAR* 4.1.

Shi, Ting et al. (2022). "Score-CAMpp: class activation map based on logarithmic transformation". In: *2022 16th IEEE International Conference on Signal Processing (ICSP)*. Vol. 1. IEEE, pp. 256–259.

Shin, Hoo-Chang et al. (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5, pp. 1285–1298.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sinha, Samarth et al. (2020). "D2rl: Deep dense architectures in reinforcement learning". In: *arXiv preprint arXiv:2010.09163*.

Snow, David A, Rens Vliegenthart, and Catherine Corrigall-Brown (2007). "Framing the French riots: A comparative study of frame variation". In: *Social forces* 86.2, pp. 385–415.

SpaceNet (Feb. 2024). *Home Page*. URL: https://spacenet.ai/challenges/ (visited on 02/16/2024).

Stow, Douglas et al. (2008). "Monitoring shrubland habitat changes through object-based change identification with airborne multispectral imagery". In: *Remote sensing of environment* 112.3, pp. 1051–1061.

Subramanya, Akshayvarun, Suraj Srinivas, and R Venkatesh Babu (2017). "Confidence estimation in deep neural networks via density modelling". In: *arXiv preprint arXiv:1707.07013*.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tahir, Arsalan et al. (2022). "Automatic target detection from satellite imagery using machine learning". In: *Sensors* 22.3, p. 1147.

Tai, Qiuqing (2014). "China's media censorship: A dynamic and diversified regime". In: *Journal of East Asian Studies* 14.2, pp. 185–210.

USGS (2022–). *USGS Landsat 9 Fact Sheet*. USGS. URL: https://pubs.usgs.gov/publication/fs20193008.

Vasu, Bhavan, Faiz Ur Rahman, and Andreas Savakis (2018). "Aerial-cam: Salient structures and textures in network class activation maps of aerial imagery". In: *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, pp. 1–5.

Voulodimos, Athanasios et al. (2018). "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018.

Vrbancic, Grega, Milan Zorman, and Vili Podgorelec (2019). "Transfer learning tuning utilizing grey wolf optimizer for identification of brain hemorrhage from head ct images". In: *StuCoSReC: proceedings of the 2019 6th student computer science research conference*, pp. 61–66.

Wang, Haofan et al. (2020). "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25.

Wei Tan, Jing et al. (2018). "Deep learning for plant species classification using leaf vein morphometric". In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.1, pp. 82–90.

Wightman, Ross, Hugo Touvron, and Hervé Jégou (2021). "Resnet strikes back: An improved training procedure in timm". In: *arXiv preprint arXiv:2110.00476*.

Wu, Congyu and Matthew S Gerber (2017). "Forecasting civil unrest using social media and protest participation theory". In: *IEEE Transactions on Computational Social Systems* 5.1, pp. 82–94.

Wu, Peng and Yumin Tan (2019). "Estimation of economic indicators using residual neural network ResNet50". In: *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 206–209.

Xie, Saining et al. (2017). "Aggregated residual transformations for deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500.

Xie, Yichun, Zongyao Sha, and Mei Yu (2008). "Remote sensing imagery in vegetation mapping: a review". In: *Journal of plant ecology* 1.1, pp. 9–23.

Yamauchi, Toshinori and Masayoshi Ishikawa (2022). "Spatial sensitive grad-cam: Visual explanations for object detection by incorporating spatial sensitivity". In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 256–260.

Zagoruyko, Sergey and Nikos Komodakis (2016). "Wide residual networks". In: *arXiv preprint arXiv:1605.07146*.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, pp. 818–833.

Zhang, Chi et al. (2019). "Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification". In: *ISPRS International Journal of Geo-Information* 8.4, p. 189.

Zhang, Liangpei, Lefei Zhang, and Bo Du (2016). "Deep learning for remote sensing data: A technical tutorial on the state of the art". In: *IEEE Geoscience and remote sensing magazine* 4.2, pp. 22–40.

Zhang, Pengbin et al. (2018). "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery". In: *Sensors* 18.11, p. 3717.

Zhou, Bolei et al. (2016). "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhou, Weiqi and Austin Troy (2008). "An object-oriented approach for analysing and characterizing urban landscape at the parcel level". In: *International Journal of Remote Sensing* 29.11, pp. 3119–3135.