

Post-Hurricane Damage Assessment Using Satellite Imagery and Geolocation Features

Quoc Dung Cao* and Youngjun Choe

Department of Industrial and Systems Engineering

University of Washington, Seattle

*Corresponding author. E-mail addresses: qcao10@uw.edu (Quoc Dung Cao) ,
ychoe@uw.edu (Youngjun Choe)

Abstract

Gaining timely and reliable situation awareness after hazard events such as a hurricane is crucial to emergency managers and first responders. One effective way to achieve that goal is through damage assessment. Recently, disaster researchers have been utilizing imagery captured through satellites or drones to quantify the number of flooded/damaged buildings. In this paper, we propose a mixed data approach, which leverages publicly available satellite imagery and geolocation features of the affected area to identify damaged buildings after a hurricane. The method demonstrated significant improvement from performing a similar task using only imagery features, based on a case study of Hurricane Harvey affecting Greater Houston area in 2017. This result opens door to a wide range of possibilities to unify the advancement in computer vision algorithms such as convolutional neural networks and traditional methods in damage assessment, for example, using flood depth or bare-earth topology. In this work, a creative choice of the geolocation features was made to provide extra information to the imagery features, but it is up to the users to decide which other features can be included to model the physical behavior of the events, depending on their domain knowledge and the type of disaster. The dataset curated in this work is made openly available (DOI: 10.17603/ds2-3cca-f398).

Summary: Creative combination of geolocation features and publicly available satellite imagery can greatly enhance hurricane damage assessment through neural network architecture.

Keywords— Disaster damage assessment, image classification, mixed data, first response, emergency management, hurricane

1 INTRODUCTION

Damage assessment after a hurricane landfall is increasingly important for emergency management as the hurricane intensity and frequency increase. The current practice of windshield survey, which relies on emergency response crews and volunteers to drive around the affected area is known to be costly and time-consuming. To speed up the process, several studies have been conducted to reduce data collection time or assist the visual inspection. One notable direction is using deep learning to detect whether a building is damaged or not after a hurricane event. In our previous work (Cao & Choe, 2020a), we have shown that using satellite imagery and annotation labels from a crowdsourcing campaign can achieve state-of-the-art performance on classifying damaged buildings based on several metrics such as accuracy, precision-recall, and F1 score. Besides satellite imagery in pure Red-Green-Blue (RGB) band, other studies have explored the use of deep learning in flood risk assessment using IKONOS-2 multispectral and panchromatic imagery (Van der Sande, De Jong, & De Roo, 2003), or time series of Landsat-5/7 satellite imagery (Skakun, Kussul, Shelestov, & Kussul, 2014). Other works have also studied to perform building damage assessment through extracting building texture feature using synthetic aperture radar (SAR) data (Q. Chen, Yang, Li, & Liu, 2019), multitemporal very high resolution SAR imagery (Pirrone, Bovolo, & Bruzzone, 2020), or multitemporal polarimetric SAR data (S.-W. Chen, Wang, & Xiao, 2018). In this work, we continue to use the high-resolution optical sensor RGB imagery since they are easier for emergency managers to interpret. Furthermore, these satellite images are increasingly available freely and quickly for disaster response thanks to public agencies (e.g. UNOOSA, NASA, USGS, NOAA) and private companies (e.g. DigitalGlobe, Planet Labs) that regularly collect satellite images for multiple purposes (S. A. Chen et al., 2018).

In fact, within the field of damage assessment, deep learning techniques have been showing promising results. A subset of deep learning models called convolutional neural network (CNN) have been applied to detect damage in concrete structures (Cha, Choi, & Büyüköztürk, 2017; Cha, Choi, Suh, Mahmoudkhani, & Büyüköztürk, 2018; Huang, Li, & Zhang, 2018), car damage (Patil, Kulkarni, Sriraman, & Karande, 2017; D. Zhang, Zhang, Li, Plummer, & Wang, 2019), or regional change detection after disaster events (Doshi, Basu, & Pang, 2018). However, these methods rely heavily on the quantity and quality of the labelled dataset, which in some cases might be unavailable or noisy. Sometimes, performance can be capped in some large image dataset such as Imagenet (Krizhevsky, Sutskever, & Hinton, 2012) and improvement in performance is marginal, regardless of model architecture.

Before the CNN era, there were established methods to assess flooding hazard risks, such as analyzing precipitation, catchment capacity, or river network analyses (Apel, Thielen, Merz, & Blöschl, 2004), generating flood outlines and depth based on topological data (Hall et al., 2003), using U.S. Army Corps of Engineers’ depth-damage curves (of Engineers, 2006), or simulating flood spreading (Gouldby, Sayers, Mulet-Marti, Hassan, & Benwell, 2008). In the seismic risk analysis field, there are also probabilistic risk assessment (Ellingwood, 2001) or defining vulnerability indices for infrastructure systems (Pitilakis, Alexoudi, Argyroudis, Monge, & Martin, 2006). These methods are still extremely valuable even in these days as a natural hazard is a natural phenomenon that obeys physical rules.

The above observation inspires us to hypothesize that there could be a potential improvement to the post-hurricane damage assessment process if we can utilize multiple types of data. We propose to utilize the optical sensor satellite imagery and other geolocation features of the individual buildings in our

damage assessment framework. This work will open up another possibility to understanding disaster damage. For example, for hurricanes, we can combine precipitation level, flooding resilience index, catchment capacity, elevation, or river networks with the imagery data to improve damage assessment and also understand which characteristics are more critical to the likelihood of damage. On the other hand, in seismic risk assessment, we can also incorporate the relative distance of the buildings/roads to the epicenter, ground shaking in the zone through various sensors, or seismic resilience index into the model, in addition to the aerial images.

In this work, we present a mixed data approach to damage assessment by utilizing the satellite imagery and other geolocation features such as building elevation and proximity to water bodies (Figure 1) to improve the performance and generalizability of our previous work (Cao & Choe, 2020a). Our contribution to the literature is two-fold. First, by considering mixed data, we can leverage more domain knowledge to understand disaster damage assessment better and boost its predictive performance. To the best of our knowledge, this work is the first one to propose a mixed data approach for combining satellite imagery and geolocation features in the disaster damage assessment literature. This novel approach holds promise for creating the synergy between ever-advancing computer vision techniques and the vast domain knowledge in disaster damage assessment. Second, as an improvement to our previous dataset in Cao and Choe (2020a), we collect the ‘*Undamaged Building*’ labels and ‘*Flooded/Damaged*’ labels from imagery of the same timestamp. Specifically, we manually build the ‘*Undamaged Building*’ labels from the undamaged region of the same imagery captured after the hurricane event. This dataset is more realistic and generalizable than the previous dataset using purely satellite imagery since it reflects the actual situation when we want to deploy this damage assessment framework

in future events. The dataset is curated and made available on the DesignSafe Data Depot (DOI: 10.17603/ds2-3cca-f398) (Cao & Choe, 2020b).

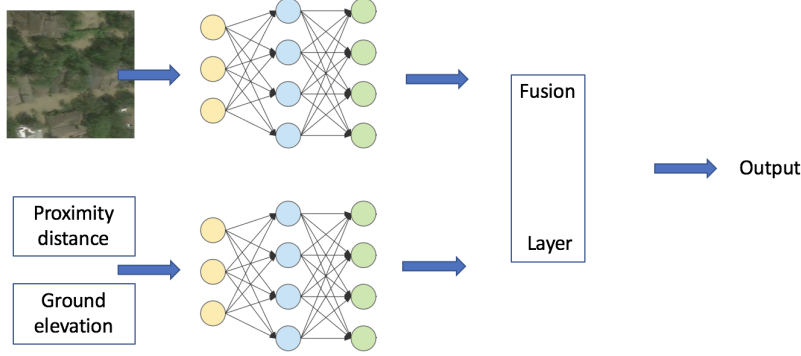


Figure 1: Workflow of the mixed data neural network model.

The remaining of this paper is organized as follows. In Section 2, we present a brief review of current literature on flood damage assessment and applications using mixed data. Section 3 describes our proposed methodology for dataset construction and model architecture. Details of the implementation and discussion of the results are presented in Section 4. Finally, Section 5 concludes this paper and draws some future research directions.

2 BACKGROUND

2.1 Relevant hurricane damage assessment practice

In this section, we review some of the relevant state-of-the-art methods in hurricane damage assessment. A more comprehensive review of the recent machine learning and deep learning methods for hurricane damage assessment is offered in Kaur, Gupta, and Singh (2021), which highlights the rapidly increasing popularity of social media and satellite imagery data for damage assessment. The use of shallow machine learning on imagery data (along with potentially other

modes of data) lost favor in the literature due to significant outperformance of deep learning. The use of multi-modal data for deep learning is yet uncommon; social media data (including images and texts) are shown conducive to multi-modal deep learning (Hao & Wang, 2020; Mouzannar, Rizk, & Awad, 2018). To the best of our knowledge, this paper presents the first multi-modal deep learning using satellite imagery data and geolocation data for damage assessment. The closest work uses the concept of multi-modal deep learning for different types of satellite imagery data (i.e., post-disaster SAR imagery and pre-disaster optical imagery) for hurricane damage assessment (Adriano et al., 2021). For earthquake damage assessment, the closest work uses optical imagery and building structural attributes for multi-modal deep learning (Miyamoto & Yamamoto, 2020).

There has always been a close relationship between flooding properties, ground topography and damage quantification. Within the field of hydrology, the role of bare-earth topography elevation is so important to hydraulic modeling of water flows that the elevation estimation methods have been studied in various works such as O’Loughlin, Paiva, Durand, Alsdorf, and Bates (2016); Yuan et al. (2019). In Kwak, Shrestha, Yorozya, and Sawano (2015), using moderate resolution imaging spectrometer (MODIS) time-series imagery, the crop damage extent at each map grid pixel (approximately 500m) is studied as a function of flood depth and flood duration. Similarly, a study by Mehrotra, Singh, Nigam, and Pal (2015) attempts to classify regions of pixels into water, vegetation, urban, and bare land at the coastal areas of Japan after an earthquake triggered a tsunami. In a separate study, relative frequency of flood inundation is shown to exhibit the same probabilistic distribution as relative water depth, which is characterized by bed elevation (Skakun et al., 2014).

The above methods mostly utilize variants of SAR imagery and/or earth

elevation from digital elevation model (DEM) in their work. SAR imagery has its own advantages in mapping surface features, or roughness pattern, and the ability to penetrate cloud cover. U.S. national agencies (e.g. NASA, USGS, NOAA) make SAR imagery available for damage assessment (Duda & Jones, 2011; Tralli, Blom, Zlotnicki, Donnellan, & Evans, 2005). In particular, NASA’s Advanced Rapid Imaging and Analysis (ARIA) project generates the Damage Proxy Map for earthquake and hurricane damage assessment based on physics-based understanding of how SAR images exhibit damages (Advanced Rapid Imaging and Analysis, 2018). However, SAR imagery could be harder for laymen (e.g. emergency managers and first responders) to interpret than optical sensor imagery. In addition, there are much fewer satellites equipped with SAR sensors than optical sensors. Our approach in this paper is to leverage the availability, partly due to the economical benefit of optical imagery versus SAR imagery, and interpretability of high-resolution RGB satellite imagery.

Our goal is to create a framework that can be readily deployed to perform hurricane damage assessment quickly in future events rather than depending on crowd-sourcing campaigns which could be time-consuming and rely on the availability and accuracy of volunteers. The framework only requires publicly available satellite imagery, such as the one used in this paper or aerial imagery collected from drones, and geographic information system (GIS) data for building coordinates and geolocation features. Although RGB satellite imagery could suffer from cloud coverage issue, which potentially delays the immediate damage assessment, there are two potential solutions to this. The first one is to leverage the current state-of-the-art cloud removal methods, such as the work in Sarukkai, Jain, Uzket, and Ermon (2020), that have demonstrated promising reconstruction accuracy in both Red-Green-Blue and infrared images. The second one is to deploy our framework repetitively over the first few days after

the event, in which SAR sensor is more expensive or infeasible to repeat the observation as frequently as the optical sensor. This is still offering practical time saving over conventional damage assessment methods such as windshield surveys (Federal Emergency Management Agency, 2016; Hristidis, Chen, Li, Luis, & Deng, 2010; Yang, Ha, Fleites, Chen, & Luis, 2011).

2.2 Deep learning with mixed data

Recent studies have demonstrated that the information is much richer when combining images and other modality of data. In Tang, Paluri, Fei-Fei, Fergus, and Bourdev (2015), image classification can be improved by including location context, derived from Global Position System (GPS) tags, of the images. Similarly, Kruk et al. (2019) also successfully learn that images on Instagram and text caption can interact with each other to inform a more complex meaning that can explain their intent, contextual, and semiotic relationships. Perhaps most related to our work are the studies showing promising results in predicting housing price using traditional housing attributes, such as area, number of rooms, zipcode, etc. combined with the house interior/exterior photo (Ahmed & Moustafa, 2016), or the neighbourhood street and aerial views (Law, Paige, & Russell, 2019). There is still a relatively smaller number of studies using multimodal data than those using images only. As highlighted by many authors in the field, there are various levels of challenges in collecting data and how to incorporate the non-image features effectively into the CNN model. In our work, we also encounter similar issues and data collection and preprocessing easily take up a major amount of work. Nonetheless, the result is really rewarding for us to achieve state-of-the-art performance in post-hurricane damage assessment. The computational cost, given the data is available, is still much more efficient than physical data collection and site survey.

3 METHODOLOGY

In this section, we describe our dataset and model architecture.

3.1 Data description

The data we used are the publicly available imagery captured after Hurricane Harvey event (post-event data), plus the coordinates annotated by crowdsourcing campaign volunteers who identified whether a building is damaged or flooded, made available by DigitalGlobe (*Tomnod*, n.d.). The raw imagery data covering the Greater Houston area was captured in about four thousand strips (~ 400 million pixels (~ 1 GB) with RGB bands per strip) in different days. In our previous work (Cao & Choe, 2020a), we used the post-event imagery to crop the images at those coordinates to build the positive labels (*Flooded/Damaged*), and pre-event data at the same coordinates to build the negative labels (*Undamaged*). This approach using temporal difference to separate the data presents some limitations in terms of modelling and usability in the future. As can be seen in Figure 2a and Figure 2b, the positive and negative labels from different timestamps may have different color scale, hue, or saturation. In addition, there is flood water almost in every positive labels, which might lead the model into water detection rather than actual damage detection, as we analyzed some of the false positive predictions in Cao and Choe (2020a).

In current work, we extract the data from the same post-event imagery (Figure 2c and Figure 2d, where it is inherently more difficult for the model to distinguish between damaged and undamaged building where flood water already invades most of the area. The color scale, hue, or saturation are also more consistent across the whole dataset to eliminate undesirable learning by the color scale.

Since the coordinates provided by DigitalGlobe only include positive labels,



(a) ‘*Undamaged Building*’ using pre-event imagery (b) ‘*Flooded/Damaged*’ using post-event imagery



(c) ‘*Undamaged Building*’ using post-event imagery (d) ‘*Flooded/Damaged*’ using post-event imagery

Figure 2: Two different ways to construct the dataset. The first way (a and b) uses different temporal information of the same location to label the data. The second way (c and d) uses different spatial information of the same timestamp to label the data. The figures, which have 46cm panchromatic resolution, represent a 256x256 pixels image cropped directly from the original strips provided by DigitalGlobe.

we need to manually collect the negative labels ourselves as shown in Figure 3. From the OpenStreetMap (OSM) API ([OpenStreetMap, n.d.](#)), we divide the area into customized, much smaller strips to extract building coordinates that do not share the same footprint with the coordinates provided by DigitalGlobe’s volunteers. We assume that the search by the volunteers are exhaustive, and

every building coordinate not found by the volunteers is considered as undamaged. This assumption is based on the fact that the crowd-sourcing campaign by TOMNOD to identify damaged buildings was designed such that multiple volunteers inspect the same images to ensure that the resulting labels are exhaustive. Despite the cross-checking effort, human errors in labeling are inevitable. This irreducible error in labeling carries over to the trained model’s performance, thereby limiting the best possible model’s ability although deep learning is relatively robust against minimal mislabeling (Khetan, Lipton, & Anandkumar, 2017; Z. Zhang & Sabuncu, 2018). The building coordinates of negative samples appearing to be in a different region from the positive samples is the result of different data sources. Using OSM, we specified different rectangular bounding boxes by the two corner coordinates to gather all the inclusive building coordinates. These boxes were chosen such that we do not include the coordinates of the positive samples. We tried our best to gather the negative samples from the region as close as possible to the positive samples but due to different systems in generating the coordinates (DigitalGlobe for ‘damaged’ and OpenStreetMap for ‘undamaged’), if the bounding box areas overlap, it is impossible to de-duplicate the coordinates.

After collecting the set of coordinates for both labels, we use Google Map Developers API (*Google Maps Platform*, n.d.) to get the elevation at these coordinates to build the elevation feature for the dataset. From the same set of coordinates, we use QGIS GRASS API to find the distance from each coordinate to their nearest water body. The raster data for Texas area water bodies are provided by the United States Geological Survey (USGS) Geographic Information System Data (*United States Geological Survey*, n.d.).

The rationale behind choosing the distance from water bodies and elevation as extra geolocation features comes from some visualization and exploratory

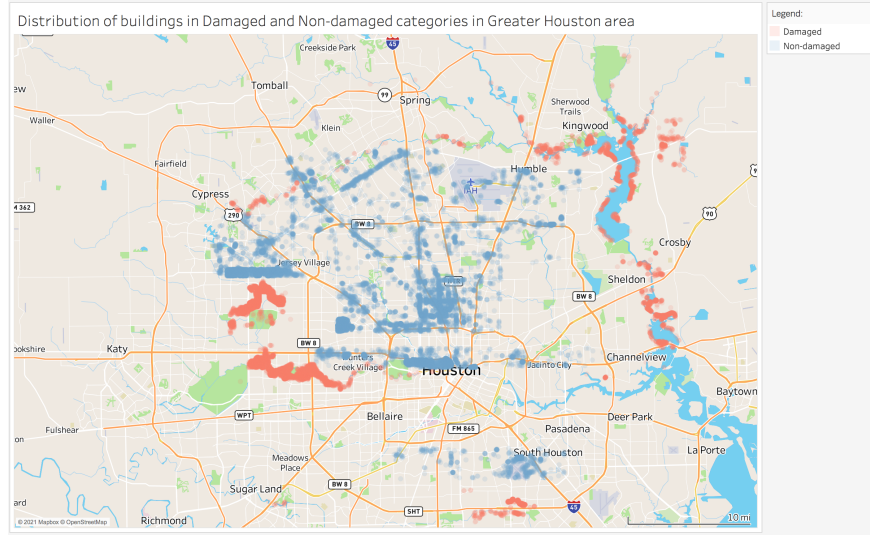
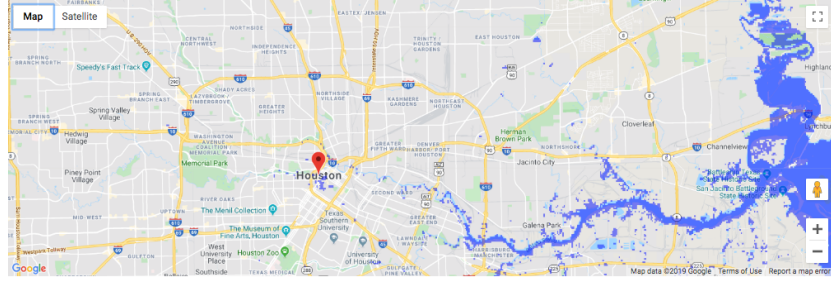


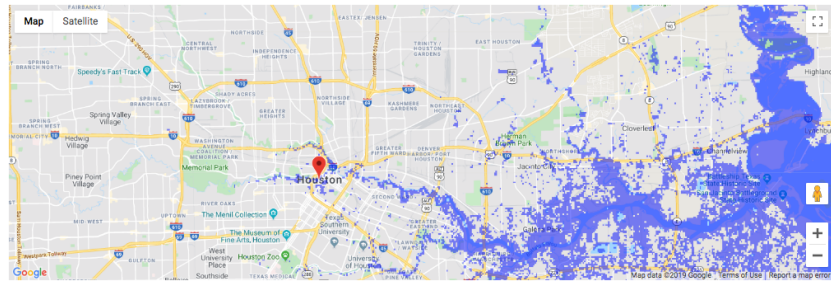
Figure 3: Post-Hurricane Harvey ‘Flooded/Damaged’ and ‘Undamaged’ building locations considered in this study.

analyses. As can be seen in Figure 3, most of the ‘Flooded/Damaged’ buildings are very close to the major water bodies in the region. This is further confirmed through a flood map simulation (*FloodMap*, n.d.) in the Houston area as shown in Figure 4 (Note that the flood extent specific for Hurricane Harvey is available online (*Hurricane Harvey Water Extent*, 2017)). As we increase the flood depth from 5 to 15 meters, areas around major water bodies have higher likelihood of being flooded. This does not mean that the distance from water body can be used as a sole signal to damage likelihood. There are more uncertainties to the probability of building damage and we hypothesize that this can be potentially an extra signal to our classification task.

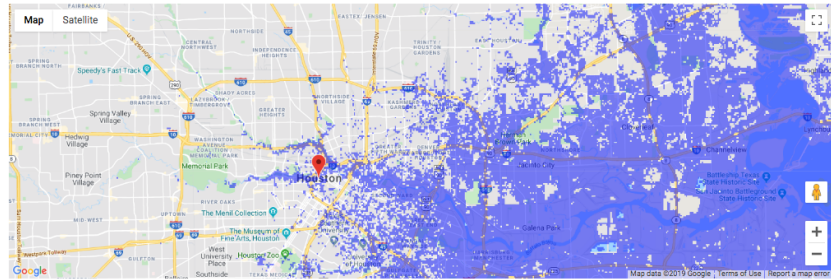
To decide the second geolocation feature, we also make some strategic comparison between individual buildings’ elevation levels and geographic coordinates. Initially, coordinates seem to be a logical choice to encode the neighbouring relationship of building cluster, which share similar flood risk factors



(a) Simulation of 5-meter flood in Houston area



(b) Simulation of 10-meter flood in Houston area



(c) Simulation of 15-meter flood in Houston area

Figure 4: Simulation of different depths (in meter) of flooding in the Houston area.

and tend to be affected together. However, elevation can be even more informative. First, it can be used as an encoder for neighbouring houses as well, as nearby houses tend to not differ much in elevation. Second, we can capture an obvious physical behaviour, in which lower elevation may result in higher likelihood of getting flooded. Last but not least, elevation may be used to generalize to other regions, whereas coordinates practically cannot. Another region may

have a different elevation, and different flood catchment capacity but the model is not expected to process absolute elevation value. We normalize the elevation to encode their relative difference within the region of interest. In future events, as long as their relative difference in elevation still prevail, we can still deploy the model trained using Houston data to quickly perform damage assessment over there.

Even with the choice of the above geolocation features, the spatial difference between the positive and negative samples does not make the classification task easier. As mentioned above, flood water is present everywhere after the hurricane event, including the Houston downtown area, in which many negative samples locate. Second, the points for negative samples are not necessarily in higher elevation. In fact, the West side of the region, which contains the majority of the positive samples, is of a much higher elevation than the East side. Furthermore, if the elevation is the sole indicator of damage likelihood, the ‘Geo Only’ model can easily outperform the mixed-data (‘Img + Geo’) model which inherits more noise from the images. These extra geolocation features, which carry some positive correlations to the damage/flood likelihood, are designed to provide extra information to the image features. As will be shown later in the experiment result, the geolocation features boost the performance of the model significantly in all classification metrics.

3.2 Model description

The models presented in Section 4 are based on deep convolutional neural network. The model consists of an image encoder for the imagery, some fully connected layers to encode the geolocation features, some fusion layers to combine the two encoded information, and a class prediction layer. An interested reader can find the full model architecture source code published at the first

author’s Github repository (Cao, 2022), in which the geolocation dataset and instruction can also be found.

For image encoder, the same convolutional setup in our previous work (Cao & Choe, 2020a) is adopted with a sequence of convolution layers, max pooling layers, followed by a fully connected layer. At the end of the image feature branch, the image encoder yields a 4-dimensional embedding vector for the imagery. For the geolocation encoder, two fully connected layers result in a 4-dimensional embedding vector. These two embedding vectors are concatenated to form a common embedding dimension of 8 in the fusion layers, which yield the final single node for class prediction after a fully connected layer of 4 nodes, as shown in Figure 5.

To benchmark against our previous work and demonstrate the additional benefits of geolocation features, we compare the performance of the proposed method with two other methods, *Img only* and *Geo only*. *Img only* is similar to our previous work, which is a pure convolutional neural network architecture that processes imagery data and output a damage prediction. *Geo only* is a multilayer perceptron network that consumes the 2 geolocation features to predict the the same response variable. Through this comparison, we show that the geolocation features are already significant features to damage prediction but with the help of extra information from the imagery feature and the mixed data architecture, we are able to achieve state-of-the-art performance.

There are some hyper-parameter tuning works in the embedding size. We would like to investigate the effect of giving the same embedding sizes to the imagery and the geolocation features. From our previous work, we know that the image encoding works quite well with the image feature alone so there is no issue with using a small embedding size (e.g 4 dimensions). The question remains whether to give the geolocation the same or smaller embedding size since

we start with only two features. This decision is informed through analyzing the performance of the model using purely imagery and geolocation features. As can be seen in Section 4, geolocation features already provide a good signal to the likelihood of building damage, almost comparable to the imagery, which leads us to give the the equal embedding size in the combined model.

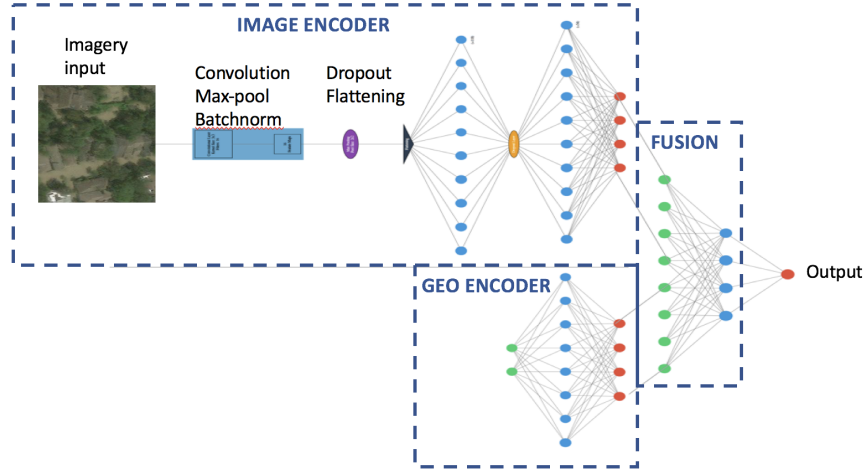


Figure 5: The mixed data neural network model that utilizes both satellite imagery and geolocation features to detect damaged building.

4 IMPLEMENTATION AND RESULT

In our previous work in Cao and Choe (2020a), we experimented with various convolutional neural network architectures, hyper-parameters tuning, and data augmentation options to pick the best set of parameters using images only (without geolocation features). Therefore, in this work, we focus on demonstrating the novelty and advantages of the mixed-data approach to the damage assessment framework. As detailed in Sec. 3.2, our architecture includes the image encoder that achieves the state-of-the-art result in Cao and Choe (2020a) and the novel geolocation encoder. This current implementation also uses a more

difficult, yet more generalizable dataset than [Cao and Choe \(2020a\)](#) as detailed in Sec. 3.1.

The network is trained to optimize with the Adam optimizer using the cross-entropy loss. We run all experiments on CentOS 7.7 (64-bit Linux), using 5 train-test splits for 70 training epochs at batch size 32. In total, we spent 25 GPU hours to run all experiments. The model is built through the *Keras* library with TensorFlow backend with a single NVIDIA K80 Tesla GPU with 64GB memory on a quad-core CPU machine.

Recall that the original imagery dataset has about four thousand strips of satellite imageries, each of which has 400 million pixels with RGB bands per strip. The individual images are cropped from the original imagery at the window size (image resolution) of 256x256 due to its better performance in the mixed data model. We experimented with two different cropping window sizes, 128x128 and 256x256 pixels, as they yielded the highest performance metrics in our previous work using images only ([Cao & Choe, 2020a](#)). The image features are augmented with random rotation, horizontal flip, vertical and horizontal shift, shear, and zoom. The geolocation features are scaled to the range of $[0, 1]$ through max-scaling to ensure the network only learn from normalized data. The binary label, *Damaged* vs *Non-Damaged*, are obtained from the TOMNOD crowdsourcing campaign.

This method relies heavily on the availability and quality of data and therefore poses some potential limitations. First, the imagery data was taken as a time series, with a lot of orthorectification and cloud coverage issues. It takes multiple iterations of visual inspection and manual processing to get a reasonable amount of usable data. Second, the geolocation data comes from different sources. Some data sources do not have complete data and/or use different formats. Preprocessing is intensive to join all the data together based on their

geo-coordinates. Nevertheless, the entire process can be done computationally and saves substantial time and manpower required for traditional damage assessment practice such as post-event windshield survey.

After cleaning and manual filtering, we are left with 13,993 positive samples (*Damaged*) and 10,384 negative samples (*Undamaged*) of unique coordinates. The dataset is split randomly to have 67% of the data as training data and 33% as test data (unseen to the training data) so that the class distribution is stratified similarly to the original train-test distribution, as well as randomly spaced across the region. The split is repeated 5 times to form 5 train-test sets, in each of which we train and test using the same model architecture to get both the mean and standard deviation of performance. Note that the train-test split is not for the purpose of hyper-parameters tuning, which was studied more extensively in our previous study, but for getting performance statistics of the method.

To study the relative contributions of two different sets of features (i.e., imagery and geolocation) to the overall model performance, we use an ablation study. Ablation studies ablate (i.e., remove) components of an AI system (e.g., modules in a deep learning model) to study the components’ relative contribution to the model’s performance (Newell, 1975; Sheikholeslami et al., 2021). In our study, the *Img only* or *Geo only* model represents a model after ablating the branch to process geolocation features or image feature, respectively, from the full model (*Img + Geo*)’s architecture (Figure 5).

We present our model performance results on the test data in Table 1 based on the probability threshold of 50% to determine a class prediction. Due to class imbalance in our dataset, the metrics used here are accuracy, F1 score, precision, and recall. Between image feature and geolocation features, the former yields better precision on average although the difference is statistically insignificant.

On the other hand, geolocation features seem to provide better recall than precision so it is more effective in detecting *Damaged* samples. This is not surprising since the geolocation features are carefully designed and we expect building damage to follow physical laws. Because of each feature’s different role, we cannot make a claim about which one is always more important. Depending on priority of the model users, probability threshold can be adjusted to trade for more recall in order to identify more *Damaged* samples. We also want to highlight the higher standard deviations of the *Img only* and *Img + Geo* models than the *Geo only* model, which is commonly observed in ultra-high parameter models such as neural networks. Nevertheless, by including the geo encoder branch, we bring down the prediction metrics standard deviation significantly, which further demonstrate the robustness of the proposed method.

Generally, combining different types of data yields quite balanced performance and improves all metrics. This ablation experiment result demonstrates the contribution of the additional geolocation features to the imagery feature because if the additional features are insignificant, they will only add further variance to the predictive model and hurt performance. The main advantage of geolocation features is their relatively greater availability in practice; collecting and processing post-hurricane satellite images depend on data vendors and thus data users (e.g., emergency managers) cannot access image features as easily as geolocation features, which are commonly available through public data sources that the users are relatively familiar with. The main advantage of image features is their visual nature that makes human interpretation intuitive. For example, the data user can assess more detailed extent of damage once the model identifies the locations of damaged buildings.

Table 1: Performance metrics across models

Method	Metrics			
	ACC	Precision	Recall	F1 score
Img only	$79.5 \pm 8.3\%$	0.88 ± 0.03	0.64 ± 0.30	0.68 ± 0.22
Geo only	$88.6 \pm 1.4\%$	0.86 ± 0.03	0.97 ± 0.003	0.91 ± 0.02
Img + Geo	$97.47 \pm 2.5\%$	0.91 ± 0.14	0.99 ± 0.003	0.94 ± 0.08

Remarks: Img only: model trained on image feature only, which corresponds to the Image Encoder box in Figure 5 followed by a class prediction output; Geo only: model trained on geolocation features only, which corresponds to the Geo Encoder box in Figure 5 followed by a class prediction output; Img + Geo: model trained on both types of data, which corresponds to the entire model in Figure 5. Each performance metric reported here shows the mean $\pm$ standard deviation across 5 train-test sets. Note that the standard error of the mean is the presented standard deviation divided by $\sqrt{5}$, thereby providing a narrower error range for the mean itself.

5 CONCLUSION

We have demonstrated that damaged buildings can be detected with 97%+ accuracy. The use of satellite imagery in RGB bands has both the benefits of availability and interpretability. In addition to the imagery feature, geolocation information can substantially improve the performance of CNN, and reduce the hyper-parameter tuning work. This paper proposed this multi-modal learning framework and illustrated it by training a model using data specific to Houston, TX after Hurricane Harvey. The framework itself holds promise to create a model that can generalize well across different regions and events as more data from more hazard events are aggregated. Developing such extensive benchmark data is a major undertaking as demonstrated in a recent international challenge (Gupta et al., 2019). Since the geolocation features used are carefully chosen as relative elevation and relative proximity to water body, the model can be adapted to deploy to other regions and events without retraining. It could be the case that the relationship between elevation and flood likelihood is specific

to regions, but we are only trying to capture the neighboring representation of the buildings through the similarity in their elevation. Therefore, to apply the model to another region, we might only need to adjust the class prediction threshold to gain more recall, if necessary. However, there is a trade-off between using too specific features to a hazard type (e.g. hurricane) such as the proximity to water bodies and generalization to other types (e.g. earthquakes). It will be helpful for the disaster management community to train a few models that have been validated on past events to be ready for deployment in the next event.

The proposed modeling framework represents one of the early steps towards automating post-hurricane damage assessment using artificial intelligence. The framework’s resulting models may supplement conventional damage assessment methods (e.g., windshield survey) by helping quickly identify locations of flooded/damaged buildings at the points of time when satellite images are collected. Because the framework utilizes the damage labels crowd-sourced from volunteers who inspect satellite images, the identifiable damages are limited to those viewable from a vertical view (e.g., visibly significant wind damage to the building roof, flood water around the building at the moment of image taken). The static nature of damage assessment and cloud coverage issue can be mitigated through repetitive image collection for the impacted area over time, but more cost-effective approaches may exist. Future work may build upon this work to get closer to the reality of automated, detailed hurricane damage assessment using multiple modes of data beyond satellite imagery and geolocation features, such as low-altitude aerial images and in-situ sensor data. Weather-related features such as wind speed and precipitation, and many other publicly available geospatial data sets (Guikema et al., 2014) hold promise to enhance the multi-modal learning framework.

In future work, we hope to investigate incorporating other features such as

catchment capacity or flood risk mapping (e.g., from a physics-based model) to further improve the performance and robustness of the model. There are two other potential directions that can be pursued following our findings in this work. The first direction is rapid, real-time damage mapping of damaged buildings. From recent efficient and instant object detection algorithms such as Redmon, Divvala, Girshick, and Farhadi (2016); Ren, He, Girshick, and Sun (2015), it would be possible to gather damage status of buildings through more accessible devices such as drones, freeing the reliance on satellite imagery. Since satellite imagery are considered to be complex to process due to their size and containing several objects at different scales (Sublime, Troya-Galvis, & Puissant, 2017), incorporating geolocation features to existing object detection algorithms can improve their metrics such as precision and recall. A potential challenge of this direction is the amount of labelled data required is usually quite large, which grows together with how complex the model is. This poses another issue of noisy or wrong labels, which leads to a second potential direction, label refinement. Recent studies have highlighted the needs of label refinement in the presence of noisy or wrong labels in large-scale datasets (Bagherinezhad, Horton, Rastegari, & Farhadi, 2018), or in remote sensing data (Shang, Lin, Jiao, & Li, 2020). From our studies, geolocation features alone already inform substantial prior knowledge about the damage likelihood. We can use that information to correct the wrong labels as necessary.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation (NSF grant CMMI-1824681, DMS-1952781, BCS-2121616). We would like to thank Maxar for data sharing through their Open Data Program.

References

- Adriano, B., Yokoya, N., Xia, J., Miura, H., Liu, W., Matsuoka, M., & Koshimura, S. (2021). Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 132–143.
- Advanced Rapid Imaging and Analysis. (2018, July). *About ARIA*. Retrieved from <https://aria.jpl.nasa.gov/about>
- Ahmed, E., & Moustafa, M. (2016). House price estimation from visual and textual features. *arXiv preprint arXiv:1609.08399*.
- Apel, H., Thieken, A. H., Merz, B., & Blöschl, G. (2004, April). Flood risk assessment and associated uncertainty. *Natural Hazards and Earth System Science*, 4(2), 295–308. Retrieved from <https://hal.archives-ouvertes.fr/hal-00299118>
- Bagherinezhad, H., Horton, M., Rastegari, M., & Farhadi, A. (2018). Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*.
- Cao, Q. D. (2022). *Post-hurricane damage assessment using satellite imagery and geolocation features*. <https://github.com/qcao10/DamageDetectionMixedData>. GitHub.
- Cao, Q. D., & Choe, Y. (2020a). Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. *Natural Hazards*, 103(3), 3357–3376.
- Cao, Q. D., & Choe, Y. (2020b). Hurricane damage assessment dataset. *Post-Hurricane Damage Assessment Using Satellite Imagery and Geolocation Features*. Retrieved from <https://doi.org/10.17603/ds2-3cca-f398>
doi: 10.17603/ds2-3cca-f398
- Cha, Y.-J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack

- damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.
- Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S., & Büyüköztürk, O. (2018). Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9), 731–747.
- Chen, Q., Yang, H., Li, L., & Liu, X. (2019). A novel statistical texture feature for sar building damage assessment in different polarization modes. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 154–165.
- Chen, S. A., Escay, A., Haberland, C., Schneider, T., Staneva, V., & Choe, Y. (2018). Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *arXiv preprint arXiv:1812.05581*.
- Chen, S.-W., Wang, X.-S., & Xiao, S.-P. (2018). Urban damage level mapping based on co-polarization coherence pattern using multitemporal polarimetric sar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8), 2657–2667.
- Doshi, J., Basu, S., & Pang, G. (2018). From satellite imagery to disaster insights. *arXiv preprint arXiv:1812.07033*.
- Duda, K. A., & Jones, B. K. (2011). USGS remote sensing coordination for the 2010 Haiti earthquake. *Photogrammetric Engineering & Remote Sensing*, 77(9), 899–907.
- Ellingwood, B. R. (2001). Earthquake risk assessment of building structures. *Reliability Engineering & System Safety*, 74(3), 251–262.
- Federal Emergency Management Agency. (2016, April). *Damage assessment operations manual*.

- Floodmap*. (n.d.). <http://www.floodmap.net/>, visited 2020-10-14.
- Google Maps Platform*. (n.d.). <https://cloud.google.com/maps-platform/maps/>, visited 2020-09-14.
- Gouldby, B., Sayers, P., Mulet-Marti, J., Hassan, M., & Benwell, D. (2008). A methodology for regional-scale flood risk assessment. In *Proceedings of the institution of civil engineers-water management* (Vol. 161, pp. 169–182).
- Guikema, S. D., Nateghi, R., Quiring, S. M., Staid, A., Reilly, A. C., & Gao, M. (2014). Predicting hurricane power outages to support storm response planning. *Ieee Access*, 2, 1364–1373.
- Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., ... Gaston, M. (2019). xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*.
- Hall, J. W., Dawson, R., Sayers, P., Rosu, C., Chatterton, J., & Deakin, R. (2003). A methodology for national-scale flood risk assessment. In *Proceedings of the institution of civil engineers-water and maritime engineering* (Vol. 156, pp. 235–247).
- Hao, H., & Wang, Y. (2020). Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction*, 51, 101760. doi: <https://doi.org/10.1016/j.ijdrr.2020.101760>
- Hristidis, V., Chen, S.-C., Li, T., Luis, S., & Deng, Y. (2010). Survey of data management and analysis in disaster situations. *Journal of Systems and Software*, 83(10), 1701–1714.
- Huang, H.-w., Li, Q.-t., & Zhang, D.-m. (2018). Deep learning based image recognition for crack and leakage defects of metro shield tunnel. *Tunnelling and Underground Space Technology*, 77, 166–176.
- Hurricane Harvey water extent*. (2017). <https://arcg.is/5DSbX>.
- Kaur, S., Gupta, S., & Singh, S. (2021). Hurricane damage detection using ma-

- chine learning and deep learning techniques: A review. In *Iop conference series: Materials science and engineering* (Vol. 1022, p. 012035).
- Khetan, A., Lipton, Z. C., & Anandkumar, A. (2017). Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc. Retrieved from <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., & Divakaran, A. (2019). Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.
- Kwak, Y., Shrestha, B. B., Yorozya, A., & Sawano, H. (2015). Rapid damage assessment of rice crop after large-scale flood in the cambodian floodplain using temporal spatial data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7), 3700–3709.
- Law, S., Paige, B., & Russell, C. (2019). Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5), 1–19.
- Mehrotra, A., Singh, K. K., Nigam, M. J., & Pal, K. (2015, May 01). Detection of tsunami-induced changes using generalized improved fuzzy radial basis function neural network. *Natural Hazards*, 77(1), 367–381. Retrieved from <https://doi.org/10.1007/s11069-015-1595-z> doi: 10.1007/s11069-015-1595-z
- Miyamoto, T., & Yamamoto, Y. (2020). Using multimodal learning model for earthquake damage detection based on optical satellite imagery and

- structural attributes. In *Igarss 2020-2020 ieee international geoscience and remote sensing symposium* (pp. 6623–6626).
- Mouzannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. In *Iscram*.
- Newell, A. (1975). A tutorial on speech understanding systems. In *Speech recognition: Invited papers presented at the 1974 IEEE symposium* (p. 43).
- of Engineers, U. S. A. C. (2006). *Depth-damage relationships for structures, contents, and vehicles and content-to-structure value ratios (csvr) in support of the donaldsonville to the gulf, louisiana, feasibility study*.
- O’Loughlin, F., Paiva, R., Durand, M., Alsdorf, D., & Bates, P. (2016). A multi-sensor approach towards a global vegetation corrected srtm dem product. *Remote Sensing of Environment*, 182, 49–59.
- OpenStreetMap. (n.d.). <https://www.openstreetmap.org/>, visited 2020-09-14.
- Patil, K., Kulkarni, M., Sriraman, A., & Karande, S. (2017). Deep learning based car damage classification. In *2017 16th ieee international conference on machine learning and applications (icmla)* (pp. 50–54).
- Pirrone, D., Bovolo, F., & Bruzzone, L. (2020). An approach to unsupervised detection of fully and partially destroyed buildings in multitemporal vhr sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 5938–5953.
- Pitilakis, K., Alexoudi, M., Argyroudis, S., Monge, O., & Martin, C. (2006). Earthquake risk assessment of lifelines. *Bulletin of Earthquake Engineering*, 4(4), 365–390.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).
- Sarukkai, V., Jain, A., Uzkent, B., & Ermon, S. (2020). Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1796–1805).
- Shang, R., Lin, J., Jiao, L., & Li, Y. (2020). Sar image segmentation using region smoothing and label correction. *Remote Sensing*, 12(5), 803.
- Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., & Dowling, J. (2021). Autoablation: Automated parallel ablation studies for deep learning. In *Proceedings of the 1st workshop on machine learning and systems* (pp. 55–61).
- Skakun, S., Kussul, N., Shelestov, A., & Kussul, O. (2014). Flood hazard and flood risk assessment using a time series of satellite images: A case study in namibia. *Risk Analysis*, 34(8), 1521–1537.
- Sublime, J., Troya-Galvis, A., & Puissant, A. (2017). Multi-scale analysis of very high resolution satellite images using unsupervised techniques. *Remote Sensing*, 9(5), 495.
- Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., & Bourdev, L. (2015). Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision* (pp. 1008–1016).
- Tomnod. (n.d.). <http://blog.maxar.com/news-events/2019/in-the-blink-of-an-eye-looking-back-on-nine-years-with-tomnod>, visited 2019-04-27.
- Tralli, D. M., Blom, R. G., Zlotnicki, V., Donnellan, A., & Evans, D. L. (2005). Satellite remote sensing of earthquake, volcano, flood, landslide

- and coastal inundation hazards. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(4), 185–198.
- United States Geological Survey. (n.d.). <https://www.usgs.gov/products/data-and-tools/gis-data/>, visited 2020-09-14.
- Van der Sande, C., De Jong, S., & De Roo, A. (2003). A segmentation and classification approach of ikonos-2 imagery for land cover mapping to assist flood risk and flood damage assessment. *International Journal of applied earth observation and geoinformation*, 4(3), 217–229.
- Yang, Y., Ha, H.-Y., Fleites, F., Chen, S.-C., & Luis, S. (2011). Hierarchical disaster image classification for situation report enhancement. In *Information reuse and integration (iri), 2011 ieee international conference on* (pp. 181–186).
- Yuan, T., Lee, H., Yu, H., Jung, H. C., Madson, A., Sheng, Y., & Beighley, E. (2019). Mapping forested floodplain topography using insar and radar altimetry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(12), 5189–5198.
- Zhang, D., Zhang, Y., Li, Q., Plummer, T., & Wang, D. (2019). Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 ieee 39th international conference on distributed computing systems (icdcs)* (pp. 1221–1232).
- Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.